Robust stimulus detection with imprecise spiking phase

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich 12
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Benjamin Straub
aus Heusenstamm

Frankfurt am Main 2018
(D30)

Vom Fachbereich 12 der

Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Andreas Bernig

Gutachter: Prof. Dr. Gaby Schneider, Prof. Dr. Anton Wakolbinger

Datum der Disputation: 10.12.2018

Summary

Precise timing of spikes between different neurons has been found to convey reliable information beyond the spike count. In contrast, the role of small phase delays with high temporal variability, as reported for example in oscillatory activity in the visual cortex, remains largely unclear. This issue becomes particularly important considering the high speed of neuronal information processing, which is assumed to be based on only a few milliseconds, or oscillation cycles within each processing step.

We investigate the role of small and imprecise phase delays with a stochastic spiking model that is strongly motivated by experimental observations. Within individual oscillation cycles the model contains only two signal parameters describing directly the rate and the phase. We specifically investigate two quantities, the probability of correct stimulus detection and the probability of correct change point detection, as a function of these signal parameters and within short periods of time such as individual oscillation cycles.

Optimal combinations of the signal parameters are derived that maximize these probabilities and enable comparison of pure rate, pure phase and combined codes. In particular, the gain in detection probability when adding imprecise phases to pure rate coding increases with the number of stimuli. More interestingly, imprecise phase delays can considerably improve the process of detecting changes in the stimulus, while also decreasing the probability of false alarms and thus, increasing robustness and speed of change point detection.

The results are applied to parameters extracted from empirical spike train recordings of neurons in the visual cortex in response to a number of visual stimuli. The results suggest that near-optimal combinations of rate and phase parameters can be implemented in the brain, and that phase parameters could particularly increase the quality of change point detection in cases of highly similar stimuli.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Prof. Dr. Gaby Schneider for her continuous support, ideas, motivation, and immense knowledge. Her guidance in neuroscience helped me throughout my research and the writing of this thesis.

My sincere thanks go to Prof. Dr. Anton Wakolbinger, who provided me an opportunity to join and support his outstanding teaching.
I gratefully acknowledge the financial support of my PhD studies from the 'Jugendstiftung', founded by Hanna Wickschrath. I would also like to thank Prof. Dr. Anton Wakolbinger and Prof. Dr. Ralph Neininger who helped provide funding for my position.

I am grateful to Prof. Dr. Danko Nikolić for an inspiring and fruitful discussion. Additionally, I thank the authors of the neurophysiological studies (Havenith et al., 2011), which motivated the thesis.

Further, I thank my colleagues and office mates Matthias Gärtner and Stefan Albert for the stimulating discussions and wonderful working atmosphere.

Finally, I would like to thank my wife Vanessa and my family, especially my brother Armin, for the constant support of my studies.

Contents

Chapter 1

Introduction

Information processing in the neural cortex is based on electronic discharges, called *spikes*, which are emitted by neurons. To investigate the response to an input, called *stimulus*, the emitted spikes are measured over a time interval $[0, \mathcal{T}]$. The output is called *spike train* and is denoted by $\mathscr{S}^{(m)} = (t_1^{(m)}, \ldots, t_{n_1}^{(m)})$ with $0 \leq t_1^{(m)} \leq \cdots \leq t_{n_m}^{(m)} \leq \mathcal{T}$, where the superscript m indicates that this spike train belongs to neuron m. So for each neuron we obtain a time series of spike responses, cf. Figure 1.1 red bars for two neurons.

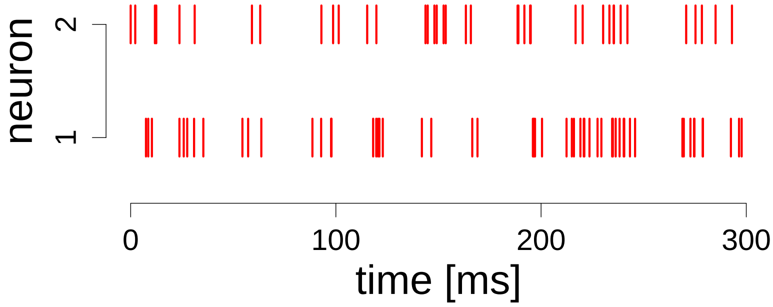


Figure 1.1: Spike trains of two parallel neurons with a recording time of $[0, 300\,\text{ms}]$. Each spike is represented as a red bar.

In experiments it was observed that dependent on the stimulus neurons tend to fire with some time delay, called *phase* delay. We are especially interested in empirical neurons as reported in Havenith et al. (2011). The authors recorded eight neurons in response to 12 stimuli, which were drifting sinusoidal gratings of which the drifting direction rotated in steps of 30°. They observed that besides the number of spikes, called *rate*, also the phase delays vary systematically with the stimulus. So for each neuron $m \in \{1, \ldots, M\}$, $M = 8$, they measured a stimulus specific rate λ_s and phase φ_s, $s \in \{1, \ldots, S\}$, $S = 12$. However, the phase delays can be measured with high precision only if we have a long recording interval $[0, \mathcal{T}]$, but cannot be identified in short time periods such as individual oscillation cycles (Schneider et al., 2006). Therefore, we call these phase delays small or imprecise. The role of such imprecise phase delays in information processing is largely unclear and it remains unclear whether and to which extent such imprecise phases can contain information in addition to the number of spikes. This question becomes particularly important considering the high speed of neuronal

information processing. As reaction times in behavioral experiments are often too short to allow long temporal averaging, information processing is assumed to take place in only a few milliseconds, or oscillation cycles within each processing step (Osram et al., 1999; Gautrais and Thorpe, 1998; Abeles, 1994).
It is the aim of the thesis to investigate whether and to which extent imprecise spiking phases can contribute to information processing within short periods of time, such as individual oscillation cycles. Furthermore, we aim at investigating parameter combinations of rate and phase that can optimize information processing, in order to enable comparison to empirical observations.

Modeling of spiking activity We use a modified version of a stochastic spike train model, the GLO (Gaussian Locking to a free Oscillator, (Bingmer, 2012)), that was able to precisely describe a variety of temporal properties related to spike timing inherent in individual spike trains (Bingmer et al., 2011; Schiemann et al., 2012) and their temporal interactions (Schneider and Nikolić, 2008). We generalize the original GLO to M neurons, which share the same background oscillation.
The GLO model is a doubly stochastic mechanism that includes the three stimulus properties rate, phase and synchronous oscillation. Figure 1.2 shows a visualization of the GLO assumptions for $M = 2$ neurons. Synchronous oscillation across multiple units can in this model for example be specified by assuming that two units share the same background oscillation (see Figure 1.2 dashed lines), which can be used to measure the degree of utilized synchrony to which they share the same background oscillation (Schneider, 2008; Schneider and Nikolić, 2006). Experimental recordings of the visual cortex of the cat have shown that the rate and phase parameters can differ across units. In particular, they can also show systematic stimulus dependence for individual neurons.
Synchronous oscillation. The GLO model assumes an unobservable background oscillation $\mathbb{B}$, which is a random walk with independent and normally distributed increments with mean $\mu_B > 0$ and variance $\sigma_B^2 \geq 0$, cf. Figure 1.2 black dashed lines, i.e.,

$$\mathbb{B} = \dots, B_{-2}, B_{-1}, B_0, B_1, B_2, \dots; \quad (B_i - B_{i-1}) \sim \mathcal{N}\left(\mu_B, \sigma_B^2\right) \forall i \in \mathbb{Z}.$$

Rate or number of spikes. Consider neuron $m \in \{1, \dots, M\}$ and stimulus $s \in \{1, \dots, S\}$, each beat B_b is assumed to give rise to an independent Poisson number of spikes $N_s^{(m)}$ with parameter $\lambda_s^{(m)} \geq 0$, cf. Figure 1.2 number of red bars.
Phase or spike timing. In the second stage, the random spike times $X_{is}^{(m)}$, $i = 1, \dots, N_s^{(m)}$, are placed independent around the beat B_b, $b \in \mathbb{Z}$, according to a normal distribution with expectation $\varphi_s^{(m)} \in \mathbb{R}$ and variance $\sigma^2 \geq 0$, cf. Figure 1.2 timing of red bars. So we assume that the precision of the spike timing is equal for all stimuli and neurons.
The spiking response of neuron m within an oscillation cycle can be described by an inhomogeneous Poisson process with intensity (see Proposition 1.1.17)

$$\rho_s^{(m)}(t) = \frac{\lambda_s^{(m)}}{\sqrt{2\pi}} \exp\left(-\frac{(\varphi_s^{(m)} - t)^2}{2}\right), s \in \{1, \dots, S\}.$$

If we talk about disjoint oscillation cycles, we assume the maximal phase parameter $\varphi_M := \max_{s,m} |\varphi_s^{(m)}|$ is small in respect to μ_B, i.e., $\mu_B \gg (\sigma_B^2 + \sigma^2) + \varphi_M$, resulting in almost no overlap between adjacent firing intensities.

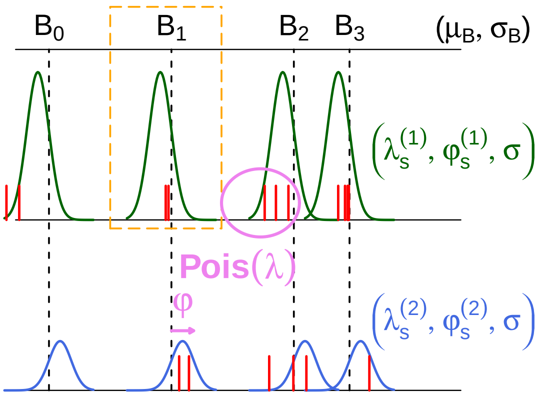

B0
B1
B2
B3
(μ_B, σ_B)
$\left(\lambda_s^{(1)}, \varphi_s^{(1)}, \sigma\right)$
Pois(λ)
φ
$\left(\lambda_s^{(2)}, \varphi_s^{(2)}, \sigma\right)$

Figure 1.2: GLO-model for two neurons responding to stimulus s. The background oscillation (dotted line) is a stationary point process with independent and normally distributed intervals with parameter (μ_B, σ_B). The red bars in the first row are the spikes generated by neuron 1, in the second row of neuron 2. The number of spikes of neuron m reacting to stimulus s at each beat B_i is $Pois(\lambda_s^{(m)})$, cf. pink circle. Every spike is placed around its birth beat according to $\mathcal{N}(\varphi_s^{(m)}, \sigma^2)$, cf. pink arrow. The corresponding firing intensity is shown in green (neuron 1) and in blue (neuron 2). Neuron 1 (2) has rate $\lambda_s^{(1)}$ $(\lambda_s^{(2)})$ and phase $\varphi_s^{(1)}$ $(\varphi_s^{(2)})$ for the presented stimulus s. Neuron 2 has lower rate, but higher phase.

Quantify information The perspective of the work is to analyze if small phases can contain additional information compared to only the number of spikes (rate code), what is commonly accepted as relevant signal component. Therefore, we focus on two different approaches:
The first one accounts for the necessity to recognize the correct stimulus out of the set of $\{1, \ldots, S\}$ stimuli, especially in a short time period, see Section 2. The basic concept is illustrated in Figure 1.3 for $S = 2$ stimuli, $M = 1$ neuron and one cycle with known background beat. We observe a sequence of spikes (red bars) and we know the theoretic firing intensity for each stimulus (blue for stimulus 1 and green for stimulus 2). With that knowledge we would decide for stimulus 1, as it is more likely to produce such a spike sequence. For this decision the number of spikes, represented by the rate λ, and the spike times, represented by the phase, are crucial. However, the information contained in the spike times decreases if we decrease the shift between the two spiking intensities. The objective is to quantify the additional information contained in such small phases dependent on the parameter ranges observed in experiments ($\lambda \in [0, 4]$ and $\varphi \in [0, 0.75]$ for $\sigma = 1$, see Appendix A) and the number of stimuli. This requires knowledge about optimal coding, separately for only using the spike numbers, only the spike times or both simultaneously. In Section 5 we compare our theoretic results to empirical neurons reported by Havenith et al. (2011).

The second approach accounts for the necessity to recognize when the stimulus changes, see Section 4. Here we assume we observe a sequence of cycles (observation as illustrated in Figure 1.3) and want to detect the change in the stimulus correctly. Therefore, we assume we start with a general idea (continuous prior information) about the rate and phase parameters and want to detect the change points, where the rate and phase parameters change according to

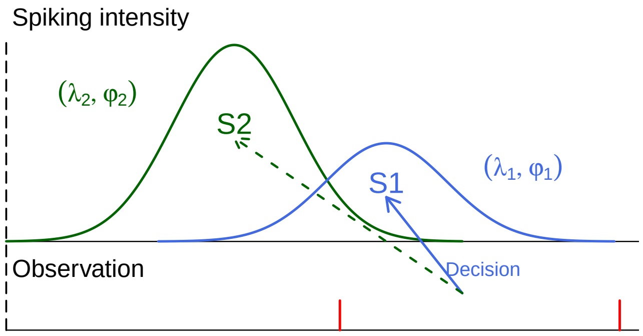


Figure 1.3: Decision task $S = 2$ stimuli and $M = 1$ neuron. The red bars illustrate the observed spikes. The theoretic firing intensity corresponding to stimulus 1 is shown in blue (stimulus 2 in green). For the presented observation we would decide for stimulus 1 (solid arrow).

the prior information. In Section 5 we consider the $M = 8$ empirical neurons of Havenith et al. (2011) corresponding to $S = 12$ stimuli, where it is appropriate to consider the explicit discrete prior information. However, in both cases our objective is to quantify, if the simultaneous analysis of rate and phase improves the change point detection compared to a pure rate analysis. Improvement is quantified by the increase in the number of correctly detected change points and by the decrease of falsely detected change points.

Overview of the thesis After a short introduction to the theory of point processes we formally introduce the GLO, see Section 1.1. In Section 1.2 we illustrate how small phase delays can be measured in empirical spike trains using the cross correlation histogram (CCH) and calculate the theoretic cross correlation function (CCF) of two GLO spike trains with the same background beat. We finish the introduction by discussing neurophysiological aspects in Section 1.3.
In Section 2 we consider the task of neurons to identify the correct stimulus out of S possible stimuli. Therefore, in Section 2.1 we focus on one neuron and investigate the probability of correct stimulus detection, p_D, first within a single oscillation cycle, as a function of rate, λ, and phase, φ (Section 2.1.1). In particular in Section 2.1.2, we investigate the maximal possible increase in this probability when including phase in addition to rate parameters. To this end, we first optimize p_D only on the basis of rate and of phase individually and then investigate p_D for the optimal combination of rate and phase parameters. Our results suggest that a rate and phase code can increase p_D compared to a pure rate code, particularly in cases with many stimuli. Second, optimal parameter combinations can be pure rate codes, pure phase codes or mixed codes, depending on the parameter range allowed for rate and phase parameters. In the case of high (also called precise) phases for example, phase coding would be preferred to rate coding. No specific correlation between the size of rate and phase parameters was observed in an optimal parameter set. In Section 2.1.3 we introduce a circular order of the stimuli, based on the empirical data, and focus on the probability to misclassify stimuli with a fixed distance. Thereby our aim of maximizing p_D shifts to minimizing the distance weighted detection error e_D. Even if this increases the computational cost, the structure of the optimal rate and phase parameters simplifies, as it is no more optimal to code a stimulus

with medium rate and medium phase. In Section 2.1.4 we compare our approach applying the Bayesian decision rule with a well known classification technique, the Linear Discriminant Analysis (LDA). Interestingly even if some assumptions of LDA are crucially violated, the results of both approaches are comparable, especially for a high number of stimuli. In Section 2.1.5 we explore the effect of observing two oscillation cycles on the optimal phase parameters and if a rate and phase code can still increase the detection probability compared to a pure rate code. Basically the optimal coding properties found in Section 2.1.2 continue to hold, but due to the additional uncertainty of the spike allocation to the correct oscillation cycle, the ability of a rate and phase code to increase the detection probability decreases compared to one cycle. Finally in Section 2.2 we generalize our procedure to M neurons, whereby we determine the detection probability in case of more than two stimuli by simulations. Here our results suggest that imprecise phases can increase the detection probability only for $S \geq 2^M$ stimuli.

In Section 3 we give a short introduction to Bayesian inference and an overview of general known results we draw on in Section 4. First we give general notations (Section 3.1) and an illustrative presentation of the Bayesian procedure, where we motivate the main theoretic results with a basic example (Section 3.2). Basic results about an appropriate prior distribution, which ensures posterior consistency, can be found in Section 3.2.4. To detect changes in the stimulus we are especially interested in the application of a Bayesian change point algorithm, where a computational efficient access to the posterior and predictive distribution is necessary. In Section 3.4 we formalize the concept of conjugacy and see that in case of an exponential family distribution a standard conjugate prior distribution exists and the predictive distribution can be determined analytically. Also the useful property of posterior linearity in the expectation of the sufficient statistic holds in general for the exponential family distribution and its standard conjugate prior.

In Section 4 we investigate the performance of pure rate and combined analyses with respect to the detection of changes in the stimulus. Thus, we analyze whether changes in the phase that may occur simultaneously to changes in the rate may improve the probability of correct change point detection, while reducing the probability of falsely detecting a change. For this purpose we extend a Bayesian change point detection algorithm (Adams and MacKay, 2007; Wilson et al., 2010) to the bivariate case and investigate its improved performance in the bivariate case over the univariate rate case when applying a newly proposed fast online decision process (Section 4.2). Our results in Section 4.3 suggest that a pure rate analysis shows a high number of falsely detected change points. In contrast, the bivariate analysis using rate and phase can considerably enhance robustness, i.e., decrease the number of falsely detected changes, while also increasing the number of correctly detected change points. In Section 4.4 we present an approach to account for special knowledge about the stimuli structure, while still providing a computationally efficient change point detection. Here we observe that a pure phase analysis can show comparable results in the change point detection as a pure rate analysis, if we have very precise prior information. Again the bivariate analysis significantly increases the number of correctly detected change points and decreases the number of falsely detected change points compared to a pure rate or phase analysis. In Section 4.5 we extend our change point model by an unknown and random spike time precision, which changes simultaneously with rate and phase. Our results suggest that a pure phase analysis can not work reliably in this setting and overestimates structural the number of change points. Taking changes in the spike precision into consideration, the trivariate analysis, compared to a pure rate analysis, can increase the number of correctly detected change point and significantly decrease the number of false

detections, especially for high spike numbers.
In Section 5 we apply our theoretical results on the detection probability and our algorithms for change point detection to a setting of empirical neurons reported in Havenith et al. (2011). The analyses of individual empirical neurons support the theoretical considerations. Concerning the detection probability, the results suggest that near-optimal parameter combinations of rate and phase do exist in empirically recorded neurons. Regarding the detection of changes in the stimulus, the bivariate analysis using rate and phase parameters can increase the number of correctly detected change points as well as increase robustness by decreasing the number of falsely detected changes in the stimulus. Further, a simultaneous consideration of multiple empirical neurons suggests that a single oscillation cycle theoretically allows the correct identification of a stimulus, except for highly similar stimuli. This holds already for rate coding alone, where the bivariate analysis shows little improvement. However, the bivariate analysis can increase the probability of change point detection particularly for similar stimuli, while also decreasing the probability of falsely detecting change points for all pairs of stimuli.
In Appendix A we explain the parameter range for rate and phase parameters extracted from experimental data. Basic definitions and properties of distributions we draw on in the thesis, are summarized in Appendix B. A collection of the most important R-Codes can be found in Appendix C to allow a reproduction of the results in this thesis.

1.1 Spike train model

In the following we give a formal introduction to the GLO. As already mentioned we slightly modify the GLO of Bingmer et al. (2011) to capture M neurons. Nevertheless, the basic properties remain and we mainly follow Bingmer (2012): Section 1.1.1 covers a short overview of the theoretic point process setting and summarizes Section 1.2 of Bingmer (2012). This allows a formal definition of the GLO using its representation as a random counting measure, see Section 1.1.2. Basic properties of the GLO are summarized from Section 2.2 of Bingmer (2012).
More details on the theory of point processes can be found in Daley and Vere-Jones (1988); Cox and Isham (1980); Thompson (1988). For a thorough introduction to random walks see Spitzer (1976), for Poisson processes see Kingman (1993) and for renewal processes see Cox (1962).

1.1.1 Basic definitions

Let E be a *complete separable metric space* (c.s.m.s) and $\mathcal{A}$ a σ-Algebra on E. The tuple $(E, \mathcal{A})$ is called a *measurable space.* Further let $\mathcal{B}(E)$ be the smallest σ-Algebra which contains the open sets of E, called a *Borel σ-Algebra*, and the elements of $\mathcal{B}(E)$ are called Borel sets. Any measure ν defined on the Borel sets is called a *Borel measure* and is boundedly finite if $\nu(A) < \infty$ for every bounded Borel set A. The support of the measure ν is defined as the set $\mathcal{R}_\nu := \{x \in E : \nu(\{x\}) > 0\}$. The space of all boundedly finite Borel measures ν on E is denoted by $\mathcal{M}_E$. The set of counting measures $\mathcal{M}_c$ consists of all boundedly finite, integer-valued measures ν defined on the Borel subsets $\mathcal{B}(E)$. Thus, $\mathcal{M}_c$ contains the subsets of simple counting measures

$$\mathcal{M}_e := \{\nu \in \mathcal{M}_c : \nu(\{x\}) \leq 1 \, \forall x \in E\}.$$

For any Borel set $A \in \mathcal{B}(E)$, we define the indicator function $\mathbb{1}_A : E \to \{0,1\}$ via

$$\mathbb{1}_A(x) := \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{else.} \end{cases}$$

Definition 1.1.1. *A point process $\mathcal{H}$ is a measurable mapping from a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ into the measurable space $(\mathcal{M}_c, \mathcal{M}_e)$. The point process is simple when $\mathbb{P}(\mathcal{H} \in \mathcal{M}_e) = 1$.*

We are interested in spike trains, which are sequences of recorded times at which a neuron fired an action potential. The recorded times can be regarded as part of a realization of a point process on the real line, thus we restrict to the measurable space $(\mathbb{R}, \mathcal{B})$ with $\mathcal{B} = \mathcal{B}(\mathbb{R})$ denoting the Borel σ-Algebra on $\mathbb{R}$.
We write a point process $\mathcal{H}$ on the line as a sequence of events

$$\cdots < T_{-1} < T_0 < 0 \leq T_1 < T_2 < \cdots ,$$

where $\{T_i\}_{i \in \mathbb{Z}}$ denotes the occurrence times. A realization h of $\mathcal{H}$ is represented by a sequence of deterministic points

$$\cdots < t_{-1} < t_0 < 0 \leq t_1 < t_2 < \cdots .$$

Alternatively, $\mathcal{H}$ is determined by its intervals together with the first occurrence time, i.e.,

$$\{W_i\}_{i \in \mathbb{Z}} \cup \{T_1\}, \text{ where } W_i := T_{i+1} - T_i \text{ for all } i \in \mathbb{Z},$$

or by the counting measure

$$\mathcal{H}(A) = \#\{i \in \mathbb{Z} : T_i \in A\} = \sum_{i \in \mathbb{Z}} \mathbb{1}_A(T_i), \quad \forall A \in \mathcal{B}.$$

Definition 1.1.2. *A spike train $\mathscr{S}$ is the restriction of the realization h of a simple point process $\mathcal{H}$ to the recording interval $[0, \mathcal{T}]$, i.e.,*

$$\mathscr{S} := (t_1, \dots, t_n) = \{t_i \in \mathcal{R}_h : i \in \mathbb{Z}\} \cap [0, \mathcal{T}]$$

with $t_1 < t_2 \cdots < t_n$ and $n = \#\{\{t_i \in \mathcal{R}_h : i \in \mathbb{Z}\} \cap [0, \mathcal{T}]\}$.

Definition 1.1.3. *Let $r(\cdot)$ be a non-negative real valued measurable function and for $a < b$ let $R(a,b) = \int_a^b r(t)dt$. Then a point process $\mathcal{H}$ is called an inhomogeneous Poisson process, if*

$$\mathbb{P}(\mathcal{H}((a_i, b_i]) = n_i, i = 1, \dots k) = \prod_{i=1}^{k} \frac{(R(a_i, b_i))^{n_i}}{n_i!} e^{-R(a_i, b_i)}$$

for $a_i < b_i < a_{i+1}$. If $r(\cdot)$ is constant, we call $\mathcal{H}$ a homogeneous Poisson process.

Remark 1.1.4. *$R(a,b)$ can be replaced by a boundedly finite Borel measure $R(\cdot)$ which is called the intensity measure.*

Definition 1.1.5. *A point process $\mathcal{H}$ is stationary when for every $r \in \mathbb{N}$ and all bounded Borel subsets $A_1, \dots, A_r$ of $\mathbb{R}$ the joint distribution of $\{\mathcal{H}(A_1 + t), \dots, \mathcal{H}(A_r + t)\}$ does not depend on $t \in (-\infty, \infty)$.*

Remark 1.1.6. *If $\mathcal{H}$ is a homogeneous Poisson process, it is stationary. In general a Poisson process is not stationary.*

Proposition 1.1.7. *Let $\mathcal{H}_1, \mathcal{H}_2 \dots$ be a countable collection of independent Poisson processes with R_n the intensity measure of $\mathcal{H}_n$ for each n. Then the superposition*

$$\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$$

is a Poisson process with intensity measure $R = \sum_{n=1}^{\infty} R_n$.

Proof. Superposition Theorem, (Kingman, 1993), p.16. □

Definition 1.1.8. *Let $X_1, X_2, \dots$ be independent identically distributed (i.i.d.) random variables on $\mathbb{R}$ and for $n \geq 0$ let $K_n = X_1 + \dots + X_n$ with $K_0 = 0$. The sequence $(K_n)_{n \in \mathbb{N}}$ is called a random walk on $\mathbb{R}$. If X_1 is nonnegative, $(K_n)_{n \in \mathbb{N}}$ is called a renewal process.*

So, every random walk is a point process with not necessarily ordered time points $K_0, K_1, K_2, \dots$. The following Remark, cf. Bingmer (2012) p.17, tells us, how to generalize a random walk to a stationary point process. This is used for the construction of the GLO spike train model.

Remark 1.1.9. *Let X_i, $i \in \mathbb{Z}$ be i.i.d random variables with $\mathbb{E}[X_1] = \mu \in (0, \infty)$ and with a distribution not concentrated on $\{0, \pm d, \pm 2d, \dots\}$, $d > 0$. We define the random walk starting at zero and evolving to the left and to the right as*

$$K_n := X_1 + \dots + X_n \quad \textit{and} \quad K_{-n} := -(X_0 + X_{-1} + \cdots, X_{-n+1}) \textit{ with } K_0 := 0.$$

The random walk represents a point process $\mathcal{H}$ on $\mathbb{R}$ by setting

$$\mathcal{H}(A) := \#\{n \in \mathbb{Z} : K_n \in A\}.$$

To obtain a stationary point process, the origin at the time axis needs to be random. Therefore we take a large boundary representing the new origin and rename the indices of the random walk by

$$\dots, K'_{-1} := K_{\tau_a - 2},\ K'_0 := K_{\tau_a - 1},\ K'_1 := K_{\tau_a},\ K'_2 := K_{\tau_a + 1}, \dots,$$

where $\tau_a := \inf\{n \in \mathbb{N} : K_n > a\}$. With that the steps of the shifted random walk are given by

$$K'_n := \lim_{a \to \infty} K_{n + \tau_a - 1} - a, \quad n \in \mathbb{Z},$$

where the existence of the limit can be seen in Woodroofe (1982), p.18. Then $\mathcal{H}(A) := \#\{i \in \mathbb{Z} : K'_i \in A\}$ is a stationary point process and $(K_i)_{i \in \mathbb{Z}}$ is called a stationary random walk.

1.1.2 The GLO

On page 2 we have already mentioned the construction of the GLO. Here we summarize the GLO model in a more compact form using its representation as a counting measure. Therefore, we modify the definition of a GLO process of Bingmer (2012) to our setting of M simultaneous neurons. Afterwards we summarize some basic properties about the GLO, see (Bingmer, 2012) Section 2.2.

We will use different simplifications of this generalized spike train model in the thesis: In Chapter 2 in Section 2.1 we analyze rate and phase parameters of one neuron in one oscillation cycle, cf. Figure 1.2 orange box, to obtain a basic understanding of optimal coding structures. Furthermore, we give a short outlook, how a sequence of oscillation cycles will influence the optimal parameter choice, cf. Section 2.1.5. In Section 2.2 we consider two neurons and one oscillation cycle and derive general effects of the number of neurons and stimuli to optimal rate and phase parameters and evaluate the increase in the detection probability by small phases for multiple neurons.
In Chapter 4 we consider one neuron and multiple oscillation cycles, where the reference time of an oscillation cycle is known as well as the assignment of each spike to its respective oscillation cycle, to analyze the ability to detect changes in the rate and phase parameter. Using experimental data, Chapter 5, we consider $M = 8$ neurons simultaneously and analysis on the one hand the ability to decide for the correct stimulus in one oscillation cycle and on the other hand detect changes in the stimulus.

Definition 1.1.10 (GLO process M neurons)**.** *Let* $\boldsymbol{\lambda}_s = \left(\lambda_s^{(1)}, \dots, \lambda_s^{(M)}\right)$ *and* $\boldsymbol{\varphi}_s = \left(\varphi_s^{(1)}, \dots, \varphi_s^{(M)}\right)$ *be rate and phase parameters with* $\lambda_s^{(m)} \geq 0$ *and* $\varphi_s^{(m)} \in \mathbb{R}$ *for all* $m = 1, \dots, M$*. The index* s *indicates that these are rate and phase parameters of neurons responding to stimulus* s*, but the index is omitted in* N *and* X*. Furthermore let* $\sigma \in [0, \infty)$ *be the spike time precision and* $\mu_B \in (0, \infty)$ *and* $\sigma_B \in [0, \infty)$ *the parameters of the background oscillation.*
Then a process $\mathcal{G} = (\mathcal{G}^{(1)}, \dots, \mathcal{G}^{(M)})$ *is a GLO process with parameters* $(\boldsymbol{\lambda}_s, \boldsymbol{\varphi}_s, \sigma, \mu_B, \sigma_B)$*, if the counting measure representation of each* $\mathcal{G}^{(m)}$*,* $m = 1, \dots, M$*, is of the form*

$$\mathcal{G}^{(m)} = \sum_{i \in \mathbb{Z}} \sum_{j=1}^{N_i^{(m)}} \mathbb{1}_{B_i + X_{i,j}^{(m)}}, \qquad \textit{with}$$

1. *a stationary walk* $\mathbb{B} = (B_i)_{i \in \mathbb{Z}}$ *with* $B_{i+1} - B_i$ *i.i.d.* $\mathcal{N}(\mu_B, \sigma_B^2)$ $\forall\, i \in \mathbb{N}$*,*

2. *spike numbers* $N_i^{(m)} \sim Pois\left(\lambda_s^{(m)}\right)$ $\forall\, i \in \mathbb{Z}$*,*

3. *spike times* $\left(X_{i,j}^{(m)}\right)$*, where* $X_{i,j}^{(m)}$ *are i.i.d.* $\mathcal{N}\left(\varphi_s^{(m)}, \sigma^2\right)$ $\forall\, i \in \mathbb{Z}$ *and* $\forall\, j \in \mathbb{N}$*,*

4. *all variables* $X_{i_1,i_2}^{(k)}$*,* $N_{i_3}^{(k)}$ *are independent* $\forall\, i_1, i_3 \in \mathbb{Z}$*,* $\forall\, i_2 \in \mathbb{N}$ *and* $\forall\, m = 1, \dots, M$ *and independent of* $(B_i)_{i \in \mathbb{Z}}$*.*

We obtain temporally ordered spike times $\dots < T_{-1}^{(m)} < T_0^{(m)} < 0 \leq T_1^{(m)} < T_2^{(m)} < \cdots$, $m \in \{1, \dots, M\}$ of the GLO process, if we choose

$$T_k^{(m)} := \begin{cases} \inf\{t \geq 0 : \mathcal{G}^{(m)}([0,t]) = k\}, & k > 0, \\ \sup\{t < 0 : \mathcal{G}^{(m)}([t,0]) = |k| + 1\}, & k \leq 0, \end{cases} \quad \forall m \in 1, \dots, M.$$

Lemma 1.1.11. $\mathbb{B}$ *is a simple point process.*

Proof. $\mathbb{B}$ is a simple point process, if $\mathbb{P}(\mathbb{B} \in \mathcal{M}_e) = 1$. This is equivalent to the condition that all beats B_i are different from each other with probability one, which holds as

$$\begin{aligned}\mathbb{P}(\{B_i \neq B_j \, \forall i \neq j\}) &= 1 - \mathbb{P}\left(\bigcup_{i \neq j} \{B_i = B_j\}\right) \\ &\geq 1 - \sum_{i \neq j} \mathbb{P}(\{B_i = B_j\}) = 1.\end{aligned}$$

□

Lemma 1.1.12. *$\mathbb{B}$ is a stationary point process.*

Proof. Follows directly from the construction of $\mathbb{B}$ in Remark 1.1.9. □

Proposition 1.1.13. *$\mathcal{G}^{(m)}$, $m \in \{1, \dots, M\}$, is a simple and stationary point process.*

Proof. See Bingmer (2012), p.31 and p.32. □

Definition 1.1.14. *$\mathcal{H}$ is a cluster process on the c.s.m.s. E_1 with center process $\mathcal{H}_c$ on the c.s.m.s. E_2 and component processes given by the family of point processes $\{\mathcal{H}(\cdot \,|\, y) : y \in E_2\}$, when for every bounded $A \in \mathcal{B}(E_1)$*

$$\mathcal{H}(A) = \int_{E_2} \mathcal{H}(A \,|\, y) \mathcal{H}_c(dy) = \sum_{y_i \in \mathcal{H}_c(\cdot)} \mathcal{H}(A \,|\, y_i) < \infty \quad a.s.$$

$\mathcal{H}$ is called an independent cluster process, if the component processes are independent.

Lemma 1.1.15. *$\mathcal{G}^{(m)}$, $m \in \{1, \dots, M\}$, is a stationary cluster process.*

Proof. We have $E_1 = E_2 = \mathbb{R}$ and $\mathcal{H}_c = \mathbb{B}$ stationary. The component processes are given by

$$\mathcal{H}(\cdot \,|\, y_i) = \sum_{j=1}^{N_i^{(m)}} \mathbb{1}_{y_i + X_{i,j}^{(m)}},$$

with $y_i \in \mathbb{B}(\cdot)$ and it holds

$$\mathcal{H}(A) = \sum_{y_i \in \mathbb{B}(\cdot)} \mathcal{H}(A \,|\, y_i) < \infty \quad a.s.$$

□

Definition 1.1.16. *Let R be a random measure on E. A point process $\mathcal{H}$ on E is a Cox process directed by R if, conditional on R, $\mathcal{H}$ is a Poisson process $\mathcal{H}(\cdot \,|\, R)$ on E with intensity measure R, cf. Definition 1.1.3 and Remark 1.1.4.*

Proposition 1.1.17. *$\mathcal{G}^{(m)}$ is a Cox process with random intensity*

$$\rho_{\mathbb{B}}(t) = \lambda_s^{(m)} \sum_{j \in \mathbb{Z}} \phi_{B_j + \varphi_s^{(m)}, \sigma^2}(t)$$

with $\phi_{\mu,\sigma^2}(\cdot)$ denoting the normal density with expectation μ and variance σ^2 and random intensity measure

$$R(A) = \int_A \rho_{\mathbb{B}}(t) dt, \quad A \in \mathcal{B}.$$

Proof. See Bingmer (2012) p.34. □

From Proposition 1.1.17 we know that the spiking response of neuron m within a fixed cycle (the beat is known) can be described by an inhomogeneous Poisson process with intensity

$$\rho_s^{(m)}(t) = \frac{\lambda_s^{(m)}}{\sqrt{2\pi}} \exp\left(-\frac{(\varphi_s^{(m)} - t)^2}{2}\right).$$

1.2 Small phase delays in empirical spike trains

To quantify the temporal correlation between two spike trains $\mathscr{S}^{(1)}$ and $\mathscr{S}^{(2)}$ measured simultaneously of two neurons, often the cross correlation histogram (CCH) is used in practice (Moore et al., 1966; Perkel et al., 1967). The CCH is an empirical estimate of the cross correlation function (CCF), which is basically an intensity which measures the occurrence of spikes per time unit in process $\mathcal{H}_2$, conditional on a spike at a particular time point in process $\mathcal{H}_1$.

Definition 1.2.1. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be stationary point processes. The cross correlation function (CCF) of $\mathcal{H}_1$ and $\mathcal{H}_2$ is defined for lags $\ell > 0$ as*

$$f(\ell) := \lim_{\Delta_1,\Delta_2 \to 0+} \frac{\mathbb{E}\left[\mathcal{H}_2(\ell, \ell+\Delta_2) \,|\, \mathcal{H}_1(-\Delta_1, 0] > 0\right]}{\Delta_2}.$$

Proposition 1.2.2. *Let $\mathcal{G} \sim GLO\big((\lambda^{(1)}, \lambda^{(2)}), (\varphi^{(1)}, \varphi^{(2)}), \sigma, \mu_B, \sigma_B\big)$, the CCF of $\mathcal{G}_1$ and $\mathcal{G}_2$ is*

$$f(\ell) = \lambda^{(2)} \sum_{i\in\mathbb{Z}} \phi_{i\mu_b + \varphi^{(2)} - \varphi^{(1)}, |i|\sigma_B^2 + 2\sigma^2}(\ell).$$

Proof. The proof is based on Proposition 4.3 in Bingmer (2012), which calculates the auto correlation function.
According to the random counting measure representation in Definition 1.1.10 we can write

$$\begin{aligned} f(\ell) &= \lim_{\Delta_1,\Delta_2 \to 0+} \frac{\mathbb{E}\left[\mathcal{G}_2(\ell, \ell+\Delta_2) \,|\, \mathcal{G}_1(-\Delta_1, 0] > 0\right]}{\Delta_2} \\ &= \lim_{\Delta_1,\Delta_2 \to 0+} \frac{\mathbb{E}\left[\sum_{i\in\mathbb{Z}} \sum_{j=1}^{N_i^{(2)}} \mathbb{1}_{B_i + X_{i,j}^{(2)}}(\ell, \ell+\Delta_2) \,|\, \mathcal{G}_1(-\Delta_1, 0] > 0\right]}{\Delta_2}. \end{aligned} \tag{1.1}$$

The event $\{\mathcal{G}_1(-\Delta_1, 0] > 0\}$ with $\Delta_1 \to 0$ basically means that there is a spike at time zero in process $\mathcal{G}_1$. We assume that this spike comes from beat B_0, otherwise we can simply rename the beats. As all $N_i^{(2)}$ are independent and independent of all $N_i^{(1)}$, of all B_j and all $X_{k_1,k_2}^{(m)}$

($\forall i, j, k_1 \in \mathbb{Z}$ and $\forall k_2 \in \mathbb{N}$) we can write Equation 1.1 as

$$\begin{aligned}
f(\ell) &= \lim_{\Delta_1,\Delta_2\to 0+} \frac{\mathbb{E}\left[N_i^{(2)}\right]\mathbb{E}\left[\sum_{i\in\mathbb{Z}} \mathbb{1}_{B_i+X_{i,j}^{(2)}}(\ell,\ell+\Delta_2)\,|\,\mathcal{G}_1(-\Delta_1,0]>0\right]}{\Delta_2} \\
&= \lim_{\Delta_1,\Delta_2\to 0+} \frac{\lambda^{(2)}\sum_{i\in\mathbb{Z}}\mathbb{E}\left[\mathbb{1}_{B_i+X_{i,j}^{(2)}}(\ell,\ell+\Delta_2)\,|\,\mathcal{G}_1(-\Delta_1,0]>0\right]}{\Delta_2} \\
&= \lim_{\Delta_1,\Delta_2\to 0+} \lambda^{(2)}\sum_{i\in\mathbb{Z}} \frac{\mathbb{P}\left(B_i+X_{i,j}^{(2)}\in(\ell,\ell+\Delta_2)\,|\,\mathcal{G}_1(-\Delta_1,0]>0\right)}{\Delta_2}. \qquad (1.2)
\end{aligned}$$

Basically Equation 1.2 is the probability that in process $\mathcal{G}_2$ a spike from Beat B_i falls into the interval $(\ell, \ell+\Delta_2)$, given a spike at time zero in process $\mathcal{G}_1$, which comes from Beat B_0. So we want to determine

$$\lim_{\Delta_1\to 0+} \mathbb{P}\left(B_i+X_{i,j}^{(2)}\in(\ell,\ell+\Delta_2)\,|\,\mathcal{G}_1(-\Delta_1,0]>0\right). \qquad (1.3)$$

We notice that

$$B_i - B_0 + X_{i,j}^{(2)} - X_{0,1}^{(1)} \sim \mathcal{N}(i\mu_B + \varphi^{(2)} - \varphi^{(1)}, |i|\sigma_B^2 + 2\sigma^2),$$

since for $i > 0$ (analog for $i < 0$)

$$B_i - B_0 = (B_i - B_{i-1}) + (B_{i-1} - B_{i-2}) + \cdots + (B_1 - B_0) \sim \mathcal{N}(i\mu_B, |i|\sigma_B^2)$$

and $X_{i,j}^{(2)} - X_{0,1}^{(1)} \sim \mathcal{N}(\varphi^{(2)} - \varphi^{(1)}, 2\sigma^2)$ are all independent. Then, cf. Formula 1.9.1 of Liemant et al. (1988), the probability 1.3 can be written in terms of $F_{\mu,\sigma}(\cdot)$ the c.d.f. of a normal distribution with mean μ and variance σ^2 as

$$F_{i\mu_B+\varphi^{(2)}-\varphi^{(1)},|i|\sigma_B^2+2\sigma^2}(\ell+\Delta_2) - F_{i\mu_B+\varphi^{(2)}-\varphi^{(1)},|i|\sigma_B^2+2\sigma^2}(\ell).$$

From Equation 1.1 we obtain

$$\begin{aligned}
f(\ell) &= \lim_{\Delta_2\to 0} \lambda^{(2)}\sum_{i\in\mathbb{Z}} \frac{F_{i\mu_B+\varphi^{(2)}-\varphi^{(1)},|i|\sigma_B^2+2\sigma^2}(\ell+\Delta_2) - F_{i\mu_B+\varphi^{(2)}-\varphi^{(1)},|i|\sigma_B^2+2\sigma^2}(\ell)}{\Delta_2} \\
&= \lambda^{(2)}\sum_{i\in\mathbb{Z}} \phi_{i\mu_B+\varphi^{(2)}-\varphi^{(1)},|i|\sigma_B^2+2\sigma^2}(\ell).
\end{aligned}$$

□

We can easily check that the CCF in case of two GLO processes has its maximum at the phase difference $\varphi^{(2)} - \varphi^{(1)}$. In GLO spike trains we can observe the phase difference by calculating the CCH. The CCH uses a discrete binning for the time axis, see Figure 1.4 A and basically counts the number of spike pairs having a particular difference according to this difference, see Figure 1.4 B and C. So the CCH obviously depends on the chosen discretization.

Let $\mathscr{S}^{(1)} = (t_1^{(1)}, \ldots, t_{n_1}^{(1)})$ and $\mathscr{S}^{(2)} = (t_1^{(2)}, \ldots, t_{n_2}^{(2)})$ denote the empirical spike times in a recording interval $[0, \mathcal{T}]$. The recorded spike times are transformed in binary time series $\tilde{\mathscr{S}}^{(1)}$ and $\tilde{\mathscr{S}}^{(2)}$ with time resolution Δ, cf. Figure 1.4 A '0' and '1', i.e.,

$$\tilde{\mathscr{S}}^{(m)}(j\Delta) = \begin{cases} 1, & \text{if at least one spike was recorded in } [j\Delta, (j+1)\Delta] \\ 0, & \text{otherwise.} \end{cases}$$

for $j = 0, 2, \ldots, \mathcal{T}/\Delta - 1$ and $m \in \{1, 2\}$. We describe how to build the CCH between the pair $\tilde{\mathscr{S}}^{(1)}$ and $\tilde{\mathscr{S}}^{(2)}$. The CCH between the pair $\tilde{\mathscr{S}}^{(2)}$ and $\tilde{\mathscr{S}}^{(1)}$ is a mirror image of the CCH between $\tilde{\mathscr{S}}^{(1)}$ and $\tilde{\mathscr{S}}^{(2)}$. First we choose a maximal delay $L > 0$ and define a set of lags

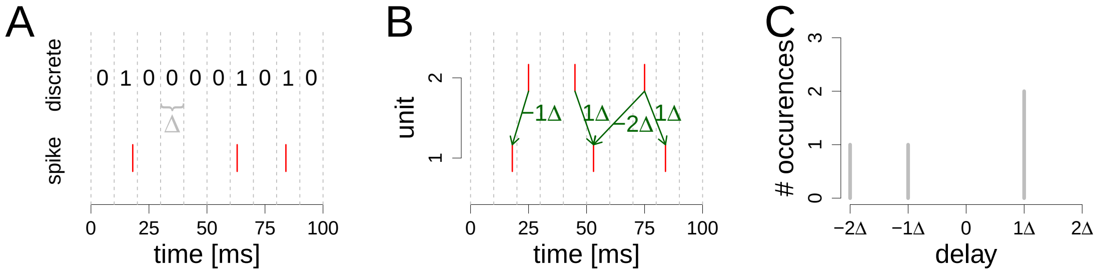


Figure 1.4: Schematic representation of a CCH. A. Recorded spike times (red) are transformed with time resolution Δ into a binary time series $(0 - 1)$. B. Schematic representation of the computation of Equation (1.4) for a maximal delay of $L = 2\Delta$. C. The resulting CCH of unit 1 and unit 2 in B.

$$\mathcal{L}_\Delta := \{j\Delta : j \in \mathbb{Z}, |j\Delta| \leq L\}.$$

Second for every $\ell \in \mathcal{L}$ the number of spikes in $\tilde{\mathscr{S}}^{(2)}$ is counted that occur ℓ time units after a spike in $\tilde{\mathscr{S}}^{(1)}$, i.e.,

$$\sum_j \tilde{\mathscr{S}}^{(1)}(j\Delta)\tilde{\mathscr{S}}^{(2)}(j\Delta + \ell). \tag{1.4}$$

For example in Figure 1.4 B all delays with $L \leq 2$ (green) are determined and the resulting CCH is shown in Figure 1.4 C. Due to the noise in the counts of a CCH and the interest in particular parameters, the counts are often smoothed with a suitable function, cf. Figure 1.5 B with raw counts in gray and in blue the counts smoothed with a Gaussian kernel. As illustration, we simulate two spike trains with a phase delay of 2 ms using the GLO. The smoothed CCH is shown in Figure 1.5 A for a maximal delay of $L = 80$ ms. If we zoom into the main peak of the CCH, cf. Figure 1.5 B, we observe the near-zero phase delay of 2 ms, green dashed line. That phase delay can be measured with high precision in case of long spike trains, here 25 s, but can be invisible in a sector of the raw spike train, cf. Figure 1.5 C, 100 ms. Experimental studies have shown that these near-zero phase delays are stimulus specific (Havenith et al., 2011), but it remains a matter of debate if these small delays can be used to improve the information processing, as the information processing is usually very fast and only few spikes can be observed. Thus, the perspective of this work is to analyze if such small phases can contain additional information compared to only the number of spikes (rate code), what is commonly accepted as relevant signal component.

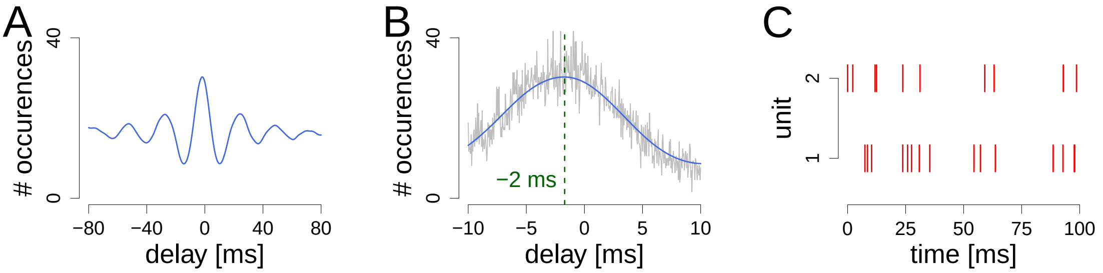


Figure 1.5: We simulate two spike trains of 1000 oscillation cycles ($\approx 25\,\mathrm{s}$) using the GLO model, described in Section 1.1. For $M = 2$ neurons we create a background oscillation of $\mu_B = 25$ and $\sigma_B = 6$. The overall spiking precision is $\sigma = 4$. The first neuron has a higher rate of $\lambda^{(1)} = 4$ expected spikes per cycle, the second neuron emits on average $\lambda^{(2)} = 2$ spikes. The phase of neuron 1 is $\varphi^{(1)} = 2$, the phase of neuron 2 is $\varphi^{(2)} = 0$. A. CCH computed for the two simulated GLO spike trains with a maximal delay $L = 80\,\mathrm{ms}$ and smoothed with a Gaussian kernel, sd=1 ms. B. Main CCH peak: In gray the raw counts, in blue the counts smoothed with a Gaussian kernel. The existing delay of 2 ms can be measured for the whole spike trains (25 s), green dashed line. C. A sector of 100 ms of the spike trains.

1.3 Neurophysiological background

In the context of neuronal coding, the identification and evaluation of different signal components has been a matter of ongoing debate. In particular, two components have been identified. First, the idea of coding by the number of action potentials, or the firing rate, dates back many decades (Adrian, 1928; Sherrington, 1933), showing for example high precision when accumulated in large populations (Softky and Koch, 1993; Shadlen and Newsome, 1998; Pouget et al., 2000). Second, also the precise timing of action potentials, called here the phase, has been found an important signal component in various cases. It has been reported across different brain structures and across sensory modalities that precise spike times can carry sensory information beyond the information contained in spike counts and can increase robustness when sensory noise is added to a stimulus (Nelken et al., 2005; Montemurro et al., 2008; Kayser et al., 2009; Cattani et al., 2015; Bieler et al., 2017).

However, neuronal firing often exhibits a high degree of variability, or noise, yielding mean phases that can be measured in the long run but not in short time scales. In the thesis we focus on one particular kind of such imprecise phases observed in synchronously oscillating neuronal populations. Synchronized oscillations are a fundamental mechanism for enabling coordinated activity (Fries, 2009; Buzsáki and Draguhn, 2004) and play a crucial role in the self-organization of developing networks (Uhlhaas. et al., 2009; Khazipov and Luhmann, 2006; Singer, 1995). The corresponding synchronized neurons have been observed to exhibit small phase delays of only a few milliseconds (Buzsáki and Chrobak, 1995; König. et al., 1995; Roelfsema et al., 1997; Schneider and Nikolić, 2006).

The role of such imprecise phase delays in information processing is largely unclear. On the one hand, these delays have been considered to reflect imprecise locking of spiking activity to the oscillation cycle (Buzsáki and Chrobak, 1995; Roelfsema et al., 1997), and they cannot be identified in short time periods such as individual oscillation cycles (Schneider et al., 2006). On the other hand, pure random variability would not explain that these phases can be measured with high precision when using accumulated data (Schneider and Nikolić, 2006), and that they can vary systematically with the stimulus (Havenith et al., 2011) or with the level of activation (Vinck et al., 2010). It therefore remains unclear whether and to which extent such imprecise phases can contain information in addition to the number of spikes. This question becomes particularly important considering the high speed of neuronal information processing. As reaction times in behavioral experiments are often too short to allow long temporal averaging, information processing is assumed to take place in only a few milliseconds, or oscillation cycles within each processing step (Osram et al., 1999; Gautrais and Thorpe, 1998; Abeles, 1994).

It is the aim of the thesis to investigate whether and to which extent imprecise spiking phases can contribute to information processing within short periods of time, such as individual oscillation cycles. Furthermore, we aim at investigating parameter combinations of rate and phase that can optimize information processing, in order to enable comparison to empirical observations.

Available approaches that investigate the information encoded in the timing of spikes often investigate situations with highly precise spike timing (Srivastava et al., 2017; Kayser et al., 2009; Nemenman et al., 2008; Thorpe et al., 2001). Theoretically, an arbitrarily large amount of information can be encoded by precise spiking in very short periods of time. In a similar way, also imprecise spike timing can convey information when averaged across large time intervals (Havenith et al., 2011; Bizley et al., 2010; Lorenzo et al., 2009; Nelken et al., 2005). However, these approaches cannot address the question as to which extent imprecise spike timing contributes to fast information processing within short time intervals. When analyzing spike timing within small temporal windows, some approaches focus on mutual information (Kayser et al., 2009; Montemurro et al., 2008, 2007). These introduce a discretized temporal binning structure and identify binary words. The difference between the estimated probability distributions on these words is then used to quantify the information. Such model free approaches have the advantage of operating on arbitrary empirical spike trains, requiring only few theoretical assumptions. However, the results crucially depend on the bin size chosen for the temporal discretization, and because potential similarities between words cannot be represented, long recordings are necessary for a good representation of the probability space, particularly for small bin sizes, which may be necessary to investigate the precise timing of spikes. In addition, quantification of the information in binary words only implicitly investigates the timing of spikes: It is possible to compare information inherent only in rate by ignoring the precise spike numbers with information inherent in rate and spike timing. However, the approach does not assign similarities to spiking patterns with similar spike timing, nor does it identify a specific parameter for the phase. It is therefore also impossible to theoretically identify optimal parameter combinations for rate and phase that are neurophysiologically plausible.

Here we present a theoretical approach that investigates the contribution of imprecise spike timing to stimulus encoding within short time intervals. In particular, we use a stochastic spiking model presented earlier (Schneider, 2008; Bingmer et al., 2011), which had been developed on a data set of spike train recordings showing small and imprecise but stimulus specific phases (Schneider and Nikolić, 2006; Schneider et al., 2006; Havenith et al., 2011). Interestingly, the model could precisely capture a variety of temporal properties related to spike timing inherent in individual spike trains (Bingmer et al., 2011; Schiemann et al., 2012) and their temporal interactions (Schneider and Nikolić, 2008).

The main part of the thesis considers only the dynamic within one oscillation cycle of the stochastic spiking model with the following assumptions. The spiking activity of a neuron within an oscillation cycle is assumed to follow an inhomogeneous Poisson process. The number of spikes within a cycle is assumed Poisson distributed with parameter λ, while the timing of every spike is independent and normally distributed with unit variance and mean (i.e., phase) φ. This simple model has the following advantages. First, it characterizes the two signal components rate and phase with two interpretable parameters and therefore allows the investigation of optimal parameter combinations. Second, it avoids the problem of choosing a bin size for temporal discretization because the phase parameter can be estimated from the mean of the raw spike times. Third, it reproduces the signal properties observed empirically, where the average phase can be measured precisely in the long run, while spike timing is highly variable across individual cycles. Fourth, it allows the theoretical investigation of the difference between pure phase, pure rate and combined rate and phase coding. To this end, we investigate two physiologically relevant quantities, namely the probability of correct stimulus detection and the probability of correct detection of a change in the stimulus. Our considerations focus on fast information processing within one or a small number of oscillation cycles. In the theoretical investigation of basic principles, we concentrate on the parameters of individual or groups of similarly activated neurons (cmp. Havenith et al., 2011; Kayser et al., 2009; Montemurro et al., 2008) throughout the thesis, which considerably reduces computational complexity. We also apply our theoretical results to parameter sets extracted from empirical recordings derived from simultaneously recorded neurons.

Chapter 2

Stimulus encoding

Here we consider the task of neurons to identify the correct stimulus out of S possible stimuli. Therefore in Section 2.1 we focus on one neuron and show in Section 2.1.1 how to determine the probability of correct stimulus detection, p_D, first within a single oscillation cycle, as a function of rate and phase parameters, cf. Lemma 2.1.3. In particular in Section 2.1.2, we investigate optimal rate and phase parameters for a given parameter range. It turns out that the parameters that maximize the detection probability always include a minimal rate parameter of zero (Lemma 2.1.2). In case of a pure rate code we show an asymptotic representation of the optimal rate parameters (Lemma 2.1.5), present an algorithmic determination of the optimal parameters in general and give the optimal parameters for the case of small rates and many stimuli (Equation 2.13). In case of a pure phase code the optimal parameters are obvious due to the symmetry of the normal distribution and equal variances. In case of a combination of rate and phase we investigate the optimal parameters numerically. Our results suggest that the phase parameter can increase p_D, particularly in cases with many stimuli. Second, optimal parameter combinations can be pure rate codes, pure phase codes or mixed codes, depending on the parameter range allowed for rate and phase parameters. In the case of precise phases for example, phase coding would be preferred to rate coding. No specific correlation between the size of rate and phase parameters was observed in an optimal parameter set.

In Section 2.1.3 we introduce a circular order of the stimuli, based on empirical data, and focus on the probability to misclassify stimuli with a fixed distance, which is closely connected to p_D (Claim 2.1.8). Thereby our aim of maximizing p_D shifts to minimizing the distance weighted detection error e_D (Definition 2.1.9). Referring to empirical data we consider $S = 12$ stimuli and compare different sets of rate and phase parameters. Even if minimizing e_D increases the computational cost, the structure of the optimal rate and phase parameters simplifies, as it is no more optimal to code a stimulus with medium rate and medium phase. This enables a natural recognition of the circular order of the stimuli in the optimal rate and and phase parameters.

In Section 2.1.4 we compare our approach applying the Bayesian decision rule with a well-known classification technique, the Linear Discriminant Analysis (LDA). First we summarize the main results about LDA, see Claim 2.1.14 and Lemma 2.1.16. Second we transfer the approach to our spiking model (Claim 2.1.20 and 2.1.22) and apply LDA to $S = 2$ and $S = 7$ stimuli. Interestingly even if some assumptions of LDA are crucially violated, the detection probability of both approaches are comparable, especially for a high number of stimuli. In case of $S = 2$ stimuli without nullstimulus even the acceptance regions are almost identical, if

we parametrize with the sum of spike times.
In Section 2.1.5 we explore the effect of two oscillation cycles on the phase and its ability to increase the detection probability. In case of a pure rate code the calculation of the detection probability can be done analogous to one cycle, as only the overall number of spikes in both cycles counts. In case of a pure phase code or a rate and phase code the calculation can be done according to Claim 2.1.25. Basically the optimal coding properties found in Section 2.1.2 continue to hold, but due to the additional uncertainty of the spike allocation to the correct oscillation cycle, the ability of the phase to increase the detection probability decreases compared to one cycle.
Finally in Section 2.2 we generalize our procedure to M neurons, whereby we determine the detection probability by simulations for more than two stimuli (Lemma 2.2.2 and Remark 2.2.4). Again we observe that p_D is maximized if both neurons have a minimal rate parameter of zero for the same stimulus (Lemma 2.2.5). Here our results suggests that two optimal neurons have a significantly higher detection probability compared to one neuron with the same overall number of spikes. However, in case of two neurons imprecise phases can increase the detection probability only for $S \geq 2^M$ stimuli.

2.1 A single neuron

Here we investigate the probability of correct stimulus detection, p_D, for a single neuron within a single oscillation cycle, as a function of the spiking parameters rate λ and phase φ. To this end we restrict to one cycle of our GLO-Model (Figure 1.2 orange box) and assume we know the start time of the cycle.
Formally, we consider a set $\{1, \dots, S\}$ of $S \in \mathbb{N}$ stimuli and rate parameters $\lambda_1, \dots, \lambda_S$, with $\lambda_s \geq 0 \, \forall s$ and phase parameters $\varphi_1, \dots, \varphi_S$, with $\varphi_s \in \mathbb{R} \, \forall s$. Note that we omit the superscript index (m) in case of one neuron. We assume that the spiking response of a neuron within an oscillation cycle is described by an inhomogeneous Poisson process with intensity (cf. Section 1.1)

$$\rho_s(t) = \frac{\lambda_s}{\sqrt{2\pi}} \exp\left(-\frac{(\varphi_s - t)^2}{2\sigma^2}\right), \quad s \in \{1, \dots, S\}. \tag{2.1}$$

This means, for stimulus s we assume a $Pois(\lambda_s)$-distributed number N_s of spikes, where the spike times X_{is}, $i = 1, \dots, N_s$ are independent and normally distributed with mean φ_s and variance 1, i.e., $\mathcal{N}(\varphi_s, 1)$. The imprecision of spikes, σ^2, can be set to 1 because only the relation of φ and σ is relevant, assuming that σ^2 is equal for all stimuli.
We derive the probability p_D to detect the correct stimulus in Section 2.1.1, assuming that all stimuli are equally likely. This probability p_D is used in Section 2.1.2 to determine the optimal rate and phase parameters that maximize the detection probability for a given parameter range and to compare the increase in the detection probability when using rate and phase to the case of pure rate or pure phase analysis.
Note that this approach investigates the theoretically optimal detection probability under the assumption that the assignment of spikes to a particular oscillation cycle is known, also implying that spiking variability is smaller than the distance between oscillation cycles. These conditions, although similar in neurophysiological recordings, do not hold perfectly in practice. However, this assumption is used here to investigate the theoretically optimal capacity of spike timing in this context.

2.1.1 The detection probability

In order to derive the probability to detect the correct stimulus among S stimuli, we assume for simplicity that all stimuli are equally likely. This assumption reduces the parameter space as the detection probability depends on the probability weights of the stimuli. However, all techniques can be applied analogously in the general case, in which detection probability will tend to increase with the inequality of stimulus weights.

We use the following notation. A realization of the random vector $B_s = (N_s, \bar{X}_s)$, where N_s denotes the number of spikes and $\bar{X}_s := \frac{1}{n}\sum_{i=1}^{n} X_{is}$ the mean spike time, is denoted by $b = (n, \bar{x}) \in \mathbb{N} \times \mathbb{R}$. For convenience, we often disregard the subscript s. Note that the mean spike time $\bar{X}$ is sufficient for the parameter φ, and n is sufficient for λ.

Given $\{N_s = n\}$ we find $\bar{X}_s \sim \mathcal{N}(\varphi_s, \sigma^2/n)$. We define for a realization $b = (n, \bar{x})$

$$P_s(b) := \begin{cases} \mathbb{P}(N_s = n)\phi_{\varphi_s,\sigma^2/n}(\bar{x}), & \text{if } n > 0, \\ \mathbb{P}(N_s = n), & \text{if } n = 0, \end{cases} \tag{2.2}$$

where $\phi_{\varphi_s,\sigma^2/n}$ denotes the density of the normal distribution with mean φ_s and variance σ^2/n at its argument. For given rate and phase parameters $\lambda_1, \ldots, \lambda_S$ and $\varphi_1, \ldots, \varphi_S$ we divide the observation space $\mathbb{N} \times \mathbb{R}$ into S acceptance regions $A_1, \ldots, A_S$ for the different stimuli that are chosen such as to maximize the *detection probability* p_D, i.e., the probability that the correct stimulus is identified,

$$p_D := \frac{1}{S}\sum_{s=1}^{S} \mathbb{P}_s(B \in A_s),$$

where $\mathbb{P}_s(B \in A_s) := \mathbb{P}(B_s \in A_s)$ for $s = 1, \ldots, S$, and $\{B_s \in A_s\}$ denotes the event that the random observation B_s falls into the acceptance region of stimulus s, i.e., stimulus s is detected.

For $S = 2$ stimuli, the optimal acceptance regions are described in Remark 2.1.1, for examples see Figure 2.1. In Lemma 2.1.3, this result is generalized to an arbitrary number S of stimuli. The idea is that according to the Bayesian decision rule (e.g., Camastra and Vinciarelli, 2015), the optimal set of acceptance regions assigns an observation b to stimulus s if $P_s(b) > P_{s'}(b)\ \forall\ s' \neq s$ under the assumption that all stimuli are equally likely. For observations b for which multiple stimuli yield the same maximal $P_s(b)$, i.e., $\exists\, \tilde{S} \subset \{1, \ldots, S\}$ with $|\tilde{S}| \geq 2$ and $P_{\tilde{s}}(b) = P_{\tilde{s}'}(b)\ \forall\, \tilde{s}, \tilde{s}' \in \tilde{S}$ and $P_{\tilde{s}}(b) > P_s(b)\ \forall \tilde{s} \in \tilde{S}, s \notin \tilde{S}$, assigning b to any of the stimuli $\tilde{s} \in \tilde{S}$ maximizes the detection probability. In the present setting, the latter case can be neglected when all phase parameters are different as it occurs with probability zero. If not all phase parameters are different, only specific and rare combinations of rate parameters can result in identical probability weight on specific observed spike numbers. In these cases, if two stimuli yield the same, maximal, probability for an observed number of spikes $n \in \mathbb{N}$, we assign this observation to the stimulus with the smaller rate parameter. If two or more stimuli have identical rate *and* phase parameters, one of these stimuli is selected uniformly.

Remark 2.1.1. *Acceptance regions for two stimuli with different phase parameters. Let (λ_1, λ_2) and (φ_1, φ_2) be rate and phase parameters for $S = 2$ stimuli and let $\varphi_1 \neq \varphi_2$. Let $N = n$ be the number of spikes and $\bar{X} = \bar{x}$ be the mean observed spike time. W.l.o.g. we only consider A_1, as A_2 can be derived analogously. The acceptance region of stimulus* 1 *is given by*

$$A_1 := \left\{ (n, \bar{x}) \,\middle|\, n \log \frac{\lambda_1}{\lambda_2} - \frac{\sqrt{n}}{\sigma}(\varphi_2 - \varphi_1)\left(\frac{\bar{x} - \varphi_1}{\sigma/\sqrt{n}} + \frac{\sqrt{n}}{\sigma}\frac{\varphi_1 - \varphi_2}{2}\right) > \lambda_1 - \lambda_2 \right\}. \tag{2.3}$$

Proof. For two stimuli with $\varphi_1 \neq \varphi_2$, acceptance region A_1 is defined by the set of all b such that $P_1(b) > P_2(b)$, or the set of all $(n, \bar{x})$ with

$$\left(\frac{\lambda_1}{\lambda_2}\right)^n e^{\lambda_2-\lambda_1} e^{-\frac{n}{2\sigma^2}\left((\bar{x}-\varphi_1)^2-(\bar{x}-\varphi_2)^2\right)} > 1.$$

Applying the natural logarithm yields the inequality in Equation 2.3.

□

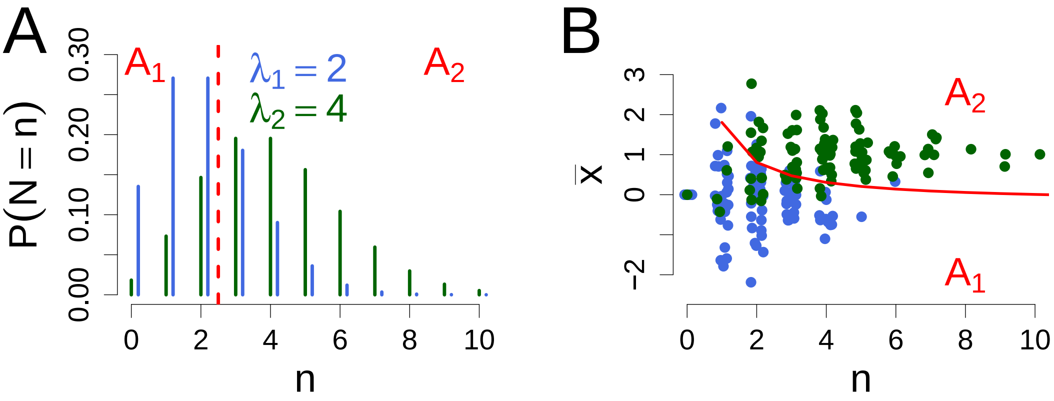


Figure 2.1: Acceptance regions A_1 and A_2 for $S = 2$ stimuli, border indicated in red. A. Phase parameters $\varphi_1 = \varphi_2$, border derived based only on the number of spikes n; $\lambda_1 = 2, \lambda_2 = 4$. B. Additional phase parameters $\varphi_1 = 0$, $\varphi_2 = 1$, acceptance regions derived on the basis of number of spikes n and mean spike time $\bar{x}$. Points indicate random realizations $(n, \bar{x})$ simulated with (λ_1, φ_1) (blue) and (λ_2, φ_2) (green).

Figure 2.1 illustrates the acceptance regions for two stimuli with different rate and equal (A) or different (B) phase parameters.
In Remark 2.1.2 we derive the detection probability for two stimuli as a function of the parameters. The general case of S stimuli, which is somewhat more technical, is given in Lemma 2.1.3..

Remark 2.1.2. *Let (λ_1, λ_2) and (φ_1, φ_2) be rate and phase parameters for $S = 2$ stimuli. From the acceptance regions (cf. Equation 2.3), the probability p_1 to correctly detect stimulus 1 is given by*

$$p_1 := \mathbb{P}_1\left(N \log\frac{\lambda_1}{\lambda_2} - \frac{\sqrt{N}}{\sigma}(\varphi_2 - \varphi_1)\left(Z + \frac{\sqrt{N}}{\sigma}\frac{\varphi_1 - \varphi_2}{2}\right) > \lambda_1 - \lambda_2\right), \tag{2.4}$$

where $N \sim Pois(\lambda_1)$ and $Z \sim \mathcal{N}(0,1)$. In detail, this probability p_1 can be written as follows.

(i) For $\varphi_1 > \varphi_2$ we obtain

$$p_1 = \sum_{n=0}^{\infty} \frac{\lambda_1^n}{n!} e^{-\lambda_1} \cdot \mathbb{P}\left(Z > \frac{\lambda_1 - \lambda_2 - n\log\frac{\lambda_1}{\lambda_2}}{\sqrt{n}\frac{\varphi_1-\varphi_2}{\sigma}} - \frac{\sqrt{n}}{2}\frac{\varphi_1 - \varphi_2}{\sigma}\right),$$

and analogously for $\varphi_1 < \varphi_2$. Again one can see that the detection probability only depends on the quotient φ/σ.

(ii) For $\varphi_1 = \varphi_2$ and $\lambda_1 < \lambda_2$, (Figure 2.1 A), we get

$$p_1 = \mathbb{P}_1 \left(N \leq \frac{\lambda_1 - \lambda_2}{\log(\lambda_1/\lambda_2)} \right),$$

which results from Equation (2.4) and the fact that if there are two stimuli with the same maximal probability weight, the stimulus with the smaller rate is assigned.

(iii) For $\lambda_1 = \lambda_2$ and $\varphi_1 = \varphi_2$, we select each stimulus with probability $1/2$.

Lemma 2.1.3. *Given S stimuli with rate parameters $\lambda_1, \ldots, \lambda_S \geq 0$ and phase parameters $\varphi_1, \ldots, \varphi_S \in \mathbb{R}$. Let*

$$G_s := \{i \in \{1, \ldots, S\} \,|\, \varphi_i = \varphi_s \wedge \lambda_i \neq \lambda_s\}$$

denote the set of stimuli with identical phase, but different rate parameter as stimulus s, and let

$$R_s := \{i \in \{1, \ldots, S\} \,|\, \varphi_i = \varphi_s \wedge \lambda_i = \lambda_s\}$$

denote the set of stimuli with identical rate and identical phase parameters as stimulus s. Let $A_1, \ldots, A_S$ denote the acceptance regions that maximize the detection probability and let $p_s := \mathbb{P}_s(B \in A_s)$ denote the probability to correctly detect stimulus s if it is present. If $\lambda_s > 0$, then p_s is given by

$$p_s = \frac{1}{|R_s|} \sum_{n=l_s}^{\lfloor u_s \rfloor} \frac{\lambda_s}{n!} e^{-\lambda_s} \cdot \mathbb{P}_s \left(\max_{\varphi_r < \varphi_s} f_{s,r}^{(n,\sigma)} < Z < \min_{\varphi_r > \varphi_s} f_{s,r}^{(n,\sigma)} \right) \tag{2.5}$$

where

$$f_{s,r}^{(n,\sigma)} := f(\lambda_s, \lambda_r, \varphi_s, \varphi_r, n, \sigma) := \frac{\lambda_r - \lambda_s - n \log \frac{\lambda_s}{\lambda_r}}{\sqrt{n} \frac{\varphi_r - \varphi_s}{\sigma}} - \frac{\sqrt{n}}{2} \frac{\varphi_s - \varphi_r}{\sigma}$$

and $Z \sim \mathcal{N}(0,1)$. The lower summation index l_s of a stimulus s with non-minimal rate, i.e., if $\exists\, r \in G_s : \lambda_r < \lambda_s$, is given by

$$l_s := \min_{k > \tilde{l}_s,\, k \in \mathbb{N}} k, \quad \textit{where} \quad \tilde{l}_s := \max_{\lambda_r < \lambda_s, r \in G_s} \left(\frac{\lambda_s - \lambda_r}{\log(\lambda_s/\lambda_r)} \right). \tag{2.6}$$

For a stimulus s with minimal rate, i.e., $\lambda_s \leq \lambda_{s'}$ $\forall s' \in G_s$ or $G_s = \emptyset$, we set $l_s = 0$ if $\forall r : \lambda_r > 0$ and $l_s = 1$ otherwise. Further

$$u_s := \min_{\lambda_r > \lambda_s, r \in G_s} \left(\frac{\lambda_s - \lambda_r}{\log(\lambda_s/\lambda_r)} \right) \quad \textit{if } \exists\, r \in G_s : \lambda_r > \lambda_s,$$

and $u_s = \infty$ otherwise. If $\lambda_s = 0$, then $p_s = 1/|R_s|$. The detection probability can be calculated as

$$p_D = \frac{1}{S} \sum_{s=1}^{S} p_s.$$

Proof. We only consider p_1, all other probabilities can be derived analogously.
In the first step we consider the case that $\lambda_s > 0\ \forall s = 1, \dots, S$ and the subcase that $\exists\, r \neq 1 : \varphi_r = \varphi_1$. In this case we can only use the number of spikes N to distinguish between such stimuli with $\varphi_r = \varphi_1$. Consider first the subset of stimuli with identical phase but different rate parameter as stimulus 1, i.e.,

$$G_1 := \{i \in \{2, \dots, S\} \,|\, \varphi_i = \varphi_1 \wedge \lambda_i \neq \lambda_1\}.$$

Using the results of Remark 2.1.2 ii), we find the acceptance regions for stimulus 1 if only stimuli of G_1 were to be distinguished as

$$N > \max_{\lambda_r \in G_1, \lambda_r < \lambda_1} \left(\frac{\lambda_1 - \lambda_r}{\log(\lambda_1/\lambda_r)} \right) =: \tilde{l}_1 \qquad \text{and} \qquad N \leq \min_{\lambda_r \in G_1, \lambda_r > \lambda_1} \left(\frac{\lambda_1 - \lambda_r}{\log(\lambda_1/\lambda_r)} \right) =: u_1.$$

Thus, if only stimuli of the set G_1 were to be distinguished, we would get

$$p_1 = \sum_{n=l_1}^{\lfloor u_1 \rfloor} \mathbb{P}_1(N = n) = \sum_{n=l_1}^{\lfloor u_1 \rfloor} \frac{\lambda_1}{n!} e^{-\lambda_1},$$

where the lower bound $l_1 := \min_{j > \tilde{l}_1,\, j \in \mathbb{N}} j$ accounts for the fact that $\tilde{l}_1$ might or might not be integer valued. If stimulus 1 has minimal rate, i.e., $\nexists\, r \in G_1 : \lambda_r < \lambda_1$, we set $l_1 = 0$ and if stimulus 1 has maximal rate, i.e., $\nexists\, r \in G_1 : \lambda_r > \lambda_1$, we set $u_1 = \infty$ and define $\lfloor \infty \rfloor := \infty$.
In the second step we still assume $\lambda_s > 0\ \forall s = 1, \dots, S$ and consider the case in which there are stimuli whose phase parameters differ from φ_1. Note that these additional stimuli can only decrease the acceptance region of stimulus 1. Therefore, it is sufficient to consider only the case $l_1 < N \leq \lfloor u_1 \rfloor$. Within this range, stimulus 1 is selected if and only if $P_1(B) > P_r(B) \quad \forall\, r \neq 1$ with $\varphi_r \neq \varphi_1$, i.e., if

$$\frac{P_1(B)}{P_r(B)} = \frac{\frac{\lambda_1^n}{n!} e^{-\lambda_1} \cdot \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \varphi_1)^2\right)}{\frac{\lambda_r^n}{n!} e^{-\lambda_r} \cdot \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \varphi_r)^2\right)} = \left(\frac{\lambda_1}{\lambda_r}\right)^n e^{\lambda_r - \lambda_1} e^{-\frac{n}{2\sigma^2}\left((\bar{x}-\varphi_1)^2 + (\bar{x}-\varphi_r)^2\right)} > 1,$$

which is equivalent to

$$n \log \frac{\lambda_1}{\lambda_r} - \frac{n(\varphi_r - \varphi_1)}{\sigma^2} \left(\bar{x} - \frac{\varphi_1 + \varphi_r}{2} \right) > \lambda_1 - \lambda_r \qquad \forall\, r \neq 1 \text{ with } \varphi_r \neq \varphi_1.$$

Combining the results of Equation (2.3) for all stimuli r with $\varphi_r < \varphi_1$ and all stimuli r with $\varphi_r > \varphi_1$ and applying the bounds l_1 and u_1, we get

$$p_1 = \sum_{n=l_1}^{\lfloor u_1 \rfloor} \frac{\lambda_1}{n!} e^{-\lambda_1} \cdot \mathbb{P}_1 \left(\max_{\varphi_r < \varphi_1} f(\lambda_1, \lambda_r, \varphi_1, \varphi_r, n, \sigma) < Z < \min_{\varphi_r > \varphi_1} f(\lambda_1, \lambda_r, \varphi_1, \varphi_r, n, \sigma) \right),$$

where $Z \sim \mathcal{N}(0, 1)$. Finally, if there are stimuli with the same φ and λ, i.e., $|R_1| > 1$, we uniformly choose one of these, which yields the factor $1/|R_1|$ in equation (2.5).
In the third step, we consider the case that there exists one stimulus s with rate $\lambda_s = 0$. This stimulus is detected if and only if $N = 0$, as

$$\mathbb{P}_s(N = 0) > \mathbb{P}_r(N = 0) \quad \text{and} \quad \mathbb{P}_s(N = i) < \mathbb{P}_r(N = i) \quad \forall\, r \neq s \text{ and } i \geq 1.$$

In that case we observe $p_s = 1/|R_s|$ because a stimulus with $\lambda_s = 0$ will always show $N = 0$. So if $\lambda_1 = 0$ we have $p_1 = 1/|R_s|$. If $\lambda_1 > 0$ we follow the arguments of the case when $\lambda_s > 0\ \forall s$, except for the subcase when stimulus 1 has minimal rate in G_1, i.e., $\nexists\, r \in G_1 : \lambda_r < \lambda_1$. In this case we set $l_1 = 1$ instead of $l_1 = 0$, which finishes the proof. □

2.1.2 Optimal parameter choices

After deriving the global detection probability p_D for given rate and phase parameters, we now investigate how these parameters should be chosen in order to maximize p_D. First, we note that p_D is not affected by a shift of the phase parameter and we can therefore assume a minimal phase parameter of zero. Then, we observe that p_D can always be increased by increasing the maximal rate $\lambda_M := \max_s \lambda_s$ and the maximal phase $\varphi_M := \max_s \varphi_s$. Therefore, we keep λ_M and φ_M fixed and derive the results as a function of these restrictions.
A further observation is that the parameters that maximize p_D also include a minimal rate parameter of zero. In the case of a pure rate code, p_D can obviously be increased by letting $\min_s \lambda_s$ decrease to zero. Also for a combined code of rate and phase parameters, with given λ_M and φ_M, p_D is maximized by letting $\min_s \lambda_s$ decrease to zero (Lemma 2.1.4).

Lemma 2.1.4. *Given S stimuli with rate parameters $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_S)$ with $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_S$ and phase parameters $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \dots, \varphi_S)$, and let $\tilde{\boldsymbol{\lambda}} = (0, \lambda_2, \dots, \lambda_S)$, i.e., λ_1 in the parameter vector $\boldsymbol{\lambda}$ is replaced by 0. Then it holds*

$$p_D(\tilde{\boldsymbol{\lambda}}, \boldsymbol{\varphi}) \geq p_D(\boldsymbol{\lambda}, \boldsymbol{\varphi}).$$

Proof. Let $A_1, \dots, A_S$ denote the disjoint acceptance regions for the stimuli $1, \dots, S$ of parameters $(\boldsymbol{\lambda}, \boldsymbol{\varphi})$, and let $\tilde{A}_1, \dots, \tilde{A}_S$ denote the acceptance regions for parameters $(\tilde{\boldsymbol{\lambda}}, \boldsymbol{\varphi})$. Recall that $A_1, \dots, A_S$ and $\tilde{A}_1, \dots, \tilde{A}_S$ are chosen such as to optimize $p_D(\boldsymbol{\lambda}, \boldsymbol{\varphi})$ and $p_D(\tilde{\boldsymbol{\lambda}}, \boldsymbol{\varphi})$, respectively. We now investigate the difference between $A_1, \dots, A_S$ and $\tilde{A}_1, \dots, \tilde{A}_S$. By setting λ_1 to zero, only the decision bounds between stimulus 1 and the other stimuli are affected, leaving the decision bounds between all other pairs of stimuli unaffected (eq. (2.3)). We can therefore divide A_1 into disjoint subsets

$$A_1 = A_{11} \cup A_{12} \cup \cdots \cup A_{1S},$$

where A_{1s} denotes the part which is allocated to region $\tilde{A}_s$, $s = 1, \dots, S$, such that

$$\tilde{A}_s = \begin{cases} A_{11} = \{0\} & \text{for } s = 1, \\ A_s \cup A_{1s} & \text{for } s = 2, \dots, S. \end{cases}$$

Let again $B = (N, \bar{X})$ denote the random vector containing the spike number and mean spike time. We split up the detection probability $p_D(\boldsymbol{\lambda}, \boldsymbol{\varphi}) = \sum_s \mathbb{P}_s(B \in A_s)$, and analogously for $p_D(\tilde{\boldsymbol{\lambda}}, \boldsymbol{\varphi})$. Then we find for stimulus 1

$$\mathbb{P}_1(\tilde{B} \in \tilde{A}_1) = 1 \geq \mathbb{P}_1(B \in A_1)$$

and for all stimuli $s = 2, \dots, k$

$$\mathbb{P}_s(\tilde{B} \in \tilde{A}_s) = \mathbb{P}_s(B \in A_s) + \mathbb{P}_s(B \in A_{1s}) \geq \mathbb{P}_s(B \in A_s).$$

□

2.1.2.1 Discrimination only on the basis of λ

Here we identify the optimal rate parameters for a given parameter range and number of stimuli for a pure rate code. To that end we assume that all phases are identical $\varphi_1 = \cdots = \varphi_S$, and $\lambda_0 = 0 \leq \lambda_1 \leq \ldots \leq \lambda_S = \lambda_M$. First, we show that the optimal rate parameters are asymptotically ($\lambda_M \to \infty$) given by a linear relation, i.e.,

$$\lambda_s = \left(\frac{s}{S}\right)^2 \lambda_M, \qquad s = 1, \ldots, S. \tag{2.7}$$

Second, to evaluate the asymptotic solution in case of medium rates, we develop an algorithmic approach that calculates the exact optimal rate parameters numerically. The algorithm makes use of the discrete structure of the Poisson distribution, using a connection between decision bounds and optimal rate parameters (Lemma 2.1.6). Already for small rates, the detection probability of the asymptotically optimal solution corresponds closely to the exact numerical solution (Figure 2.2 A).

Third, we consider the case where the maximal rate is small relative to the number of stimuli, i.e., $\lambda_M \leq S$, and show (Lemma 2.1.7) that the detection probability is maximized by $\lambda_0 = 0$, $\lambda_S = \lambda_M$ and

$$\lambda_s = \begin{cases} s, & \text{if } s < \lambda_M, \\ 0, & \text{else}, \end{cases} \qquad \text{for } s = 1, \ldots, S-1.$$

Asymptotic solution

To derive the asymptotic solution we recall that in case of identical phase parameters $\varphi_1 = \cdots = \varphi_S$, the detection probability simplifies to (cf. Remark 2.1.2 and Lemma 2.1.3)

$$p_D(\boldsymbol{\lambda}, \boldsymbol{\varphi}) = \frac{1}{S+1}\left(1 + \sum_{s=1}^{S-1} \mathbb{P}_s(\ell_{s-1} < N \leq \ell_s) + \mathbb{P}_S(N > \ell_{S-1})\right),$$

where the decision bound between stimulus s and $s+1$ is given by

$$\ell_s := \frac{\lambda_{s+1} - \lambda_s}{\log\left(\lambda_{s+1}/\lambda_s\right)} \quad \text{for } s = 1, \ldots, S-1, \quad \ell_0 := 0.$$

If λ_M is asymptotically large, also the rate parameters $\lambda_1, \ldots, \lambda_S$ are large (as they are all positive and $\lambda_0 = 0$), such that each Poisson distribution with parameter λ_s can be approximated by a normal distribution with mean λ_s and variance λ_s. In Lemma 2.1.5, we use this property to show relation (2.7). Using the asymptotic distribution requires new asymptotically optimal decision bounds and detection probability, which are denoted by $\tilde{\ell}$ and $\tilde{p}_D$, respectively.

Lemma 2.1.5. *Consider rate parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)$ with $\lambda_0 = 0 < \lambda_1 < \cdots < \lambda_S \leq \lambda_M$, and let $Z_1, \ldots, Z_S$ be independent with $Z_s \sim \mathcal{N}(\lambda_s, \lambda_s)$ for all $s = 1, \ldots, S$. Let $Z_0 = 0$ a.s. (the nullstimulus). Accordingly, $A_0 = \{0\}$ and $\mathbb{P}_0(Z_0 \in A_0) = 1$. First, for given $\boldsymbol{\lambda}$, the probability*

$$\tilde{p}_D(\boldsymbol{\lambda}, \tilde{\boldsymbol{\ell}}) = \frac{1}{S+1}\left(1 + \sum_{s=1}^{S-1} \mathbb{P}(\tilde{\ell}_{s-1} < Z_s \leq \tilde{\ell}_s) + \mathbb{P}(Z_S > \tilde{\ell}_{S-1})\right),$$

with $\tilde{\ell}_0 := 0$ and $\tilde{\boldsymbol{\ell}} = (\tilde{\ell}_1, \ldots, \tilde{\ell}_{S-1})$, is maximized by decision bounds

$$\tilde{\ell}_s^*(\boldsymbol{\lambda}) = \sqrt{\frac{\log(\lambda_{s+1}) - \log(\lambda_s) + \lambda_{s+1} - \lambda_s}{1/\lambda_s - 1/\lambda_{s+1}}} \quad \textit{for } s = 1, \ldots, S-1. \tag{2.8}$$

Second, if $\boldsymbol{\lambda}$ is not given, the vector of rates $\boldsymbol{\lambda}^* := (\lambda_1^*, \ldots, \lambda_S^*)$ which maximizes the detection probability is asymptotically given by

$$\boldsymbol{\lambda}^* := \arg\max_{\boldsymbol{\lambda}} \tilde{p}_D(\boldsymbol{\lambda}, \tilde{\boldsymbol{\ell}}^*(\boldsymbol{\lambda})) \to \left(\left(\frac{1}{S}\right)^2 \lambda_M, \left(\frac{2}{S}\right)^2 \lambda_M, \ldots, \left(\frac{S}{S}\right)^2 \lambda_M \right), \tag{2.9}$$

as $\lambda_M \to \infty$.

Proof. Equation (2.8) results by solving $f_{\lambda_s}(\tilde{\ell}_s) = f_{\lambda_{s+1}}(\tilde{\ell}_s)$, where $f_{\lambda_s}(\cdot)$ denotes the density of the $\mathcal{N}(\lambda_s, \lambda_s)$-distribution, i.e., solving

$$\frac{1}{\sqrt{2\pi\lambda_s}} \exp\left(-\frac{(\tilde{\ell}_s - \lambda_s)^2}{2\lambda_s}\right) = \frac{1}{\sqrt{2\pi\lambda_{s+1}}} \exp\left(-\frac{(\tilde{\ell}_s - \lambda_{s+1})^2}{2\lambda_{s+1}}\right)$$

for $\tilde{\ell}_s$.

In order to show equation (2.9), note that $\lambda_M \to \infty$ implies $\lambda_s^* \to \infty$ and also $(\lambda_s^* - \lambda_{s-1}^*) \to \infty$ for $s = 1, \ldots, S$, because this reduces the probability of $Z_s < 0$, $s = 1, \ldots, S$ and the overlap of adjacent densities f_{λ_s} and $f_{\lambda_{s+1}}$. Asymptotically, we thus approximate the relation between the optimal decision bounds $\tilde{\ell}_s^*$ and the optimal rate parameters λ_s^*

$$\frac{\tilde{\ell}_s^*}{\sqrt{\lambda_s^* \lambda_{s+1}^*}} = \frac{\sqrt{\frac{\log(\lambda_{s+1}^*) - \log(\lambda_s^*) + \lambda_{s+1}^* - \lambda_s^*}{\frac{\lambda_{s+1}^* - \lambda_s^*}{\lambda_s^* \lambda_{s+1}^*}}}}{\sqrt{\lambda_s^* \lambda_{s+1}^*}} \longrightarrow \frac{\sqrt{\frac{1}{1/(\lambda_s^* \lambda_{s+1}^*)}}}{\sqrt{\lambda_s^* \lambda_{s+1}^*}} = 1. \tag{2.10}$$

For the case $S = 2$, we have $\lambda_1^* = c_1 \lambda_2^*$ and recall $\tilde{\ell}_0 = 0$ for the nullstimulus. Due to the symmetry of the normal distribution, the optimal λ_1^* must be the mean $\lambda_1^* = 0.5(\tilde{\ell}_0 + \tilde{\ell}_1^*) = \tilde{\ell}_1^*/2$ between the two decision bounds. Using the asymptotic representation of $\tilde{\ell}_1^*$ (2.10), we require

$$\frac{2\lambda_1^*}{\sqrt{\lambda_1^* \lambda_2^*}} = 2\sqrt{c_1} \longrightarrow 1,$$

which yields $c_1 \to 1/4 = (1/S)^2$.

The induction from S to $S+1$ stimuli uses the fact that for a given optimal λ_S^*, the optimal relation for all λ_s^* with $s \leq S$ persists, which is denoted by $c_s = \lambda_s^*/\lambda_{s+1}^*$ $\forall\, s = 1, \ldots, S-1$. Thus it is sufficient to determine the optimal relation between λ_S^* and λ_{S+1}^*.

Again due to the symmetry of the normal distribution it holds $\lambda_S^* = \left(\tilde{\ell}_{S-1}^* + \tilde{\ell}_S^*\right)/2$, which yields

$$\sqrt{c_S} = \frac{\lambda_S^*}{\sqrt{\lambda_S^* \lambda_{S+1}^*}} = \frac{\tilde{\ell}_{S-1}^* + \tilde{\ell}_S^*}{2\sqrt{\lambda_S^* \lambda_{S+1}^*}} = \frac{1}{2}\left(\frac{\tilde{\ell}_{S-1}^*}{\sqrt{\lambda_{S-1}^* \lambda_S^*}} \sqrt{c_{S-1} c_S} + \frac{\tilde{\ell}_S^*}{\sqrt{\lambda_S^* \lambda_{S+1}^*}}\right).$$

Solving the expression for $\sqrt{c_S}$ and applying the induction $c_{S-1} \to \left(\frac{S-1}{S}\right)^2$ and (2.10) yields

$$\sqrt{c_S} = \frac{\tilde{\ell}_S^* / \sqrt{\lambda_S^* \lambda_{S+1}^*}}{2 - (\sqrt{c_{S-1}} \tilde{\ell}_{S-1}^* / \sqrt{\lambda_{S-1}^* \lambda_S^*})} \longrightarrow \frac{1}{2 - \frac{S-1}{S}} = \frac{S}{S+1}.$$

Overall the induction yields

$$\frac{\lambda_s^*}{\lambda_{S+1}^*} = \frac{\lambda_s^*}{\lambda_S^*} c_S \to \left(\frac{s}{S}\right)^2 \left(\frac{S}{S+1}\right)^2 = \left(\frac{s}{S+1}\right)^2.$$

□

Algorithmic determination

The basic idea of the algorithmic determination is to derive the optimal decision bounds instead of the optimal rate parameters (Lemma 2.1.6).

Lemma 2.1.6. *For fixed decision bounds* $L = \{0, \ell_1, \dots, \ell_{S-1}\}$ *with* $0 < \ell_1 < \ell_2 < \dots < \ell_{S-1}$, $\ell_i \in \mathbb{N}$, *and fixed* $\lambda_S = \lambda_M$ *the optimal rate parameters are*

$$\lambda_s = \left(\frac{\ell_s!}{\ell_{s-1}!}\right)^{1/(\ell_s - \ell_{s-1})}, \quad s = 1, \dots, S-1. \tag{2.11}$$

Proof. Stimulus s is detected if $N \in [\ell_{s-1}+1, \ell_s]$. We therefore need to maximize the probability

$$p_s^{\ell_{s-1},\ell_s}(\lambda_s) := \sum_{i=\ell_{s-1}+1}^{\ell_s} \frac{\lambda_s^i}{i!} e^{-\lambda_s}.$$

The derivative with respect to λ_s is

$$\frac{\partial p_s^{\ell_s,\ell_{s-1}}(\lambda_s)}{\partial \lambda_s} = \sum_{i=\ell_{s-1}+1}^{\ell_s} \left(\frac{i\lambda_s^{i-1}}{i!} e^{-\lambda_s} - \frac{\lambda_s^i}{i!} e^{-\lambda_s}\right) = e^{-\lambda_s} \lambda_s^{\ell_{s-1}} \left(\frac{1}{\ell_{s-1}!} - \frac{\lambda_s^{\ell_s - \ell_{s-1}}}{\ell_s!}\right),$$

which vanishes for the term given in Equation (2.11). As the gradient changes from positive to negative and there are no extremes at the margins, this is a maximum. □

We now derive the set of decision bounds L. For $S = 2$, $L = \{0, \ell_1\}$, with $\ell_1 \in \mathbb{N}$ because $N \in \mathbb{N}$. Recall that the optimal λ_1 is a function of ℓ_1. Therefore, we only need to determine the optimal ℓ_1 as a function of λ_2. To that end, we determine the value $\tilde{\lambda}_2$ of λ_2 for which $p_D^{(\ell)} = p_D^{(\ell+1)}$ because for reasons of monotonicity, the optimal $\ell_1 \geq \ell + 1$ for $\lambda_2 > \tilde{\lambda}_2$, and the optimal $\ell_1 \leq \ell$ for $\lambda_2 < \tilde{\lambda}_2$. The respective solution for ℓ_1 can be derived numerically (Figure 2.2 C).

In case of $S > 2$ stimuli, the optimal set of bounds L can be derived with dynamic programming, by determining the optimal lower decision bounds as a function of higher decision bounds, starting with the smallest bounds. Knowing the optimal combinations of the decision bounds, we only need to determine for which value $\tilde{\lambda}_S$ of λ_S the decision bound with the next higher ℓ_{S-1} is chosen. The optimal rate parameters are then given in Lemma 2.1.6 and are illustrated for $S = 5$ in Figure 2.2 B.

A note on the case of small rates and many stimuli, i.e., when $\lambda_M \leq S$

While the asymptotic case treated above considers the case of large λ_M, the case of small λ_M will be particularly important in the data analysis in Section 5. Therefore, we discuss this case in more detail here. In this case, one combination of rate parameters that maximizes detection probability is given by $\lambda_S = \lambda_M$, $\lambda_s = s$ for $s < \lfloor \lambda_M \rfloor$ and $\lambda_s = 0$ else. Thus, the rate of 0 is taken several times, implying that the corresponding stimuli cannot be distinguished. The main idea is that the discreteness of the Poisson distribution allows for only $\lambda_M + 1$ different decision areas if $\lambda_M \in \mathbb{N}$, given by

$$A_0 = \{0\}, A_1 = \{1\}, \dots, A_{\lambda_M} = \{\lambda_M, \lambda_M + 1, \dots\} \tag{2.12}$$

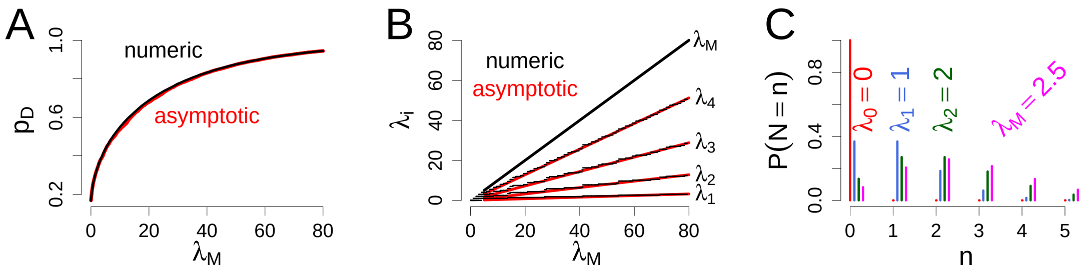


Figure 2.2: A and B. Maximal detection probability and optimal rate parameters for $S = 5$ stimuli. A. Detection probability p_D derived with the optimal numeric solution (Lemma 2.1.6) (black line). The red curve indicates the approximate asymptotic solution according to Lemma 2.1.5. B. Optimal rate parameters derived asymptotically according to Lemma 2.1.5 (straight lines) and derived numerically according to Lemma 2.1.6 (black step functions). C. Optimal rate parameters for case with $S \geq \lambda_M$, i.e., $S = 3$ stimuli and $\lambda_M = 2.5$.

and for $\lfloor \lambda_M \rfloor + 2$ areas if $\lambda_M \notin \mathbb{N}$ given by

$$A_0 = \{0\}, A_1 = \{1\}, \ldots, A_{\lfloor \lambda_M \rfloor} = \lfloor \lambda_M \rfloor, A_{\lfloor \lambda_M \rfloor + 1} = \{\lfloor \lambda_M \rfloor + 1, \lfloor \lambda_M \rfloor + 2, \ldots\} \text{ (Fig. 2.2 C).} \tag{2.13}$$

Optimal separation between these areas is obtained by choosing the rate parameters identical to the decision areas for $s \leq \lambda_M$. Thus, for $\lambda_M \notin \mathbb{N}$ and $S = \lfloor \lambda_M \rfloor + 1$ stimuli, the optimal rate parameters are $\{0, 1, 2, \ldots, \lfloor \lambda_M \rfloor, \lambda_M\}$ (where $S = \lfloor \lambda_M \rfloor$ and $\lfloor \lambda_M \rfloor = \lambda_M$ in case of λ_M integer). This is because for a fixed $j \in \mathbb{N}$ the rate parameter λ_0 that maximizes the Poisson weight on j is $\lambda_0 = j$, which maximizes the weights on each decision area. For a proof see Lemma 2.1.7.

If the number of stimuli is larger than λ_M, the choice of the rate parameters in the additional stimuli does not affect the detection probability. This is because an additional stimulus must necessarily share an already defined acceptance region with a given stimulus, while a gain in detection probability for the additional stimulus corresponds to a loss of equal size in detection probability for a given stimulus, and a stimulus with a non-integer rate parameter will never be detected.

Lemma 2.1.7. *Let $N_\lambda \sim Pois(\lambda)$ with $\lambda \geq 0$ and $N_j \sim Pois(j)$ for a fixed $j \in \mathbb{N}$. Then*

$$\mathbb{P}(N_j = j) \geq \mathbb{P}(N_\lambda = j).$$

Proof. For $\lambda := j + \epsilon$, $\epsilon \in \mathbb{R}$, we show

$$\frac{(j)^j}{k!} e^{-j} \geq \frac{(j+\epsilon)^j}{j!} e^{-(j+\epsilon)}, \qquad \text{which is equivalent to} \qquad \left(\frac{j}{j+\epsilon}\right)^j \geq e^{-\epsilon}. \tag{2.14}$$

For $\epsilon = 0$, equality holds. For $\epsilon \neq 0$ we use the bounds of the logarithm

$$1 - \frac{1}{x} < \log(x) < x - 1 \qquad \forall\, x > 1. \tag{2.15}$$

If $\epsilon < 0$, (2.14) follows with the first part of (2.15), as

$$\left(\frac{j}{j+\epsilon}\right)^j = \exp\left(j\log\left(\frac{j}{j+\epsilon}\right)\right) \overset{(2.15)}{>} \exp\left(j\left(1-\frac{j+\epsilon}{j}\right)\right) = \exp\left(-j\frac{\epsilon}{j}\right) = e^{-\epsilon}.$$

If $\epsilon > 0$, (2.14) follows with the second part of (2.15), as

$$\left(\frac{j}{j+\epsilon}\right)^j = \exp\left(-j\log\left(\frac{j+\epsilon}{j}\right)\right) \overset{(2.15)}{>} \exp\left(-j\cdot\frac{\epsilon}{j}\right) = e^{-\epsilon}.$$

□

2.1.2.2 Discrimination only on the basis of φ

Here we consider a pure phase code and identify the optimal phase parameters for a given parameter range and number of stimuli. Note that for reasons of comparability we always assume a nullstimulus with rate zero, for which we decide if no spike is observed, where the phase is naturally irrelevant. The notion of 'phase code' therefore refers here to the situation in which the parameters of all stimuli *except the nullstimulus* may differ only in the phase parameter and have the same, positive rates, i.e., $0 < \lambda_1 = \cdots = \lambda_S$ and $0 \leq \varphi_1 \leq \varphi_2 \leq \cdots \leq \varphi_S$.
Searching for the optimal vector of phase parameters corresponds to positioning the means of S normal distributions with the same variance on the interval $[0, \varphi_M]$ with minimal overlap. Due to the symmetry of the normal distribution, the phase parameters need to be chosen equidistantly in order to maximize the detection probability, i.e.,

$$\varphi_2 - \varphi_1 = \varphi_3 - \varphi_2 = \varphi_4 - \varphi_3 = \cdots = \varphi_S - \varphi_{S-1},$$

which yields

$$\varphi_s = \frac{s-1}{S-1}\cdot\varphi_M, \qquad s = 1, \ldots, S.$$

This solution holds for any fixed spike number $N = n$, where the variances of the respective normal distributions scale with $1/n$. Thus the solution holds also for a random number of spikes $N \sim Pois(\lambda)$.

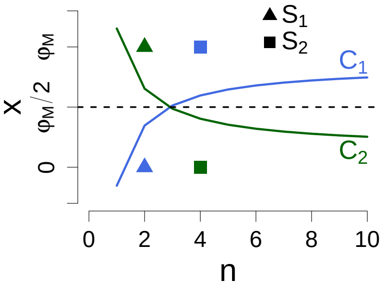
2.1.2.3 Optimal combination of λ and φ

Here we investigate how to combine rate and phase parameters in order to optimize the detection probability. To that end we consider the discrimination based simultaneously on λ and φ. W.l.o.g. we assume $\lambda_0 = 0 \leq \lambda_1 \leq, \ldots \leq \lambda_S = \lambda_M$, we set the minimum phase to zero and investigate the relations as a function of λ_M and φ_M. As the values of λ_M and φ_M crucially determine the detection probability, we focus on biologically plausible values of $\lambda_M \approx 4$ and $\varphi_M \approx 0.75$ derived in correspondence with the rate and phase parameters given in (Schneider and Nikolić, 2008) (see Materials and Methods Section A).
Assignment of rate to phase parameters under optimally: For symmetry reasons, no particular assignment between rate and phase parameters can be considered optimal. For $S = 2$, two parameter cases need to be considered, where C_1 assigns the minimal phase to the minimal rate, while C_2 assigns the minimal phase to the maximal rate (Figure 2.3). Both cases have identical detection probability, which can be seen as follows. We focus on the case $\lambda_1 > 0$

because otherwise the value of the phase parameter is irrelevant. Rearranging the acceptance regions in Remark 2.1.1 yields that in C_1 we decide for stimulus 1 if

$$\bar{x} < \frac{\lambda_1 - \lambda_2 - n \log \frac{\lambda_1}{\lambda_2}}{\frac{n}{\sigma^2}(\varphi_1 - \varphi_2)} + \frac{\varphi_1 + \varphi_2}{2},$$

and vice versa for C_2. Thus the borders of the acceptance regions of the two cases are symmetric to $(\varphi_1 + \varphi_2)/2 = \varphi_M/2$. Due to the symmetry of the normal distribution, both cases result in the same detection probability.
Also for more than two stimuli, the optimal parameter vectors do not show specific relations between the rate and phase parameters, i.e., high rates are combined both with large and small phases (Figure 2.4 C).



		C_1	C_2
Stimulus	λ	φ	φ
0	0	–	–
1	λ_1	0	φ_M
2	λ_M	φ_M	0

Figure 2.3: Two combinations of rate and phase parameters for two stimuli (stimulus 1, triangle, and stimulus 2, square). In case C_1 (blue), the smaller rate is combined with the smaller phase, and vise versa in C_2 (green). Borders of acceptance regions indicated by colored curves, which are symmetric to $\varphi_M/2$ (black dotted line).

Phase, rate and combined coding

Pure rate codes, pure phase codes or combined codes can be optimal depending on the allowed parameter ranges, i.e., on the values of φ_M and λ_M. This is illustrated in Figure 2.4 A for the case of two stimuli with $\lambda_2 = \lambda_M = 4$, $\varphi_1 = 0$, $\varphi_2 = \varphi_M$ (i.e., the case C_1). The blue curve indicates the optimal value of λ_1 determined numerically as a function of the maximal phase φ_M, where the grey lines indicate the detection probability surface. For $\varphi_M = 0$, the rate parameters are used for stimulus detection, and we find $\lambda_0 < \lambda_1 < \lambda_2$, which we call a pure rate code. With increasing φ_M, the optimal value of λ_1 increases, and we find $\lambda_1 < \lambda_2$ and $\varphi_1 < \varphi_2$, which we call a combined code. For large φ_M, we find $\lambda_1 = \lambda_M$, which we call a pure phase code. The analogous evolution can be found for $S = 7$ stimuli (Figure 2.4 C). Starting with a pure rate code for $\varphi_M = 0$, more and more stimuli are coded with maximal rate by increasing φ_M, ending up in a pure phase-code.
Increase in detection probability by adding phase parameters: Interestingly, even small phases can increase the detection probability compared to a pure rate code. The increase in the optimal detection probability as a function of φ_M is depicted in Figure 2.4 B and D for two and seven stimuli, respectively. Already for two stimuli, a maximum phase of $\varphi_M = 0.5$ results in a higher detection probability of a phase code over a pure rate code. For more stimuli, the combined use of rate and phase parameters can considerably increase the detection probability (Figure 2.4 D).

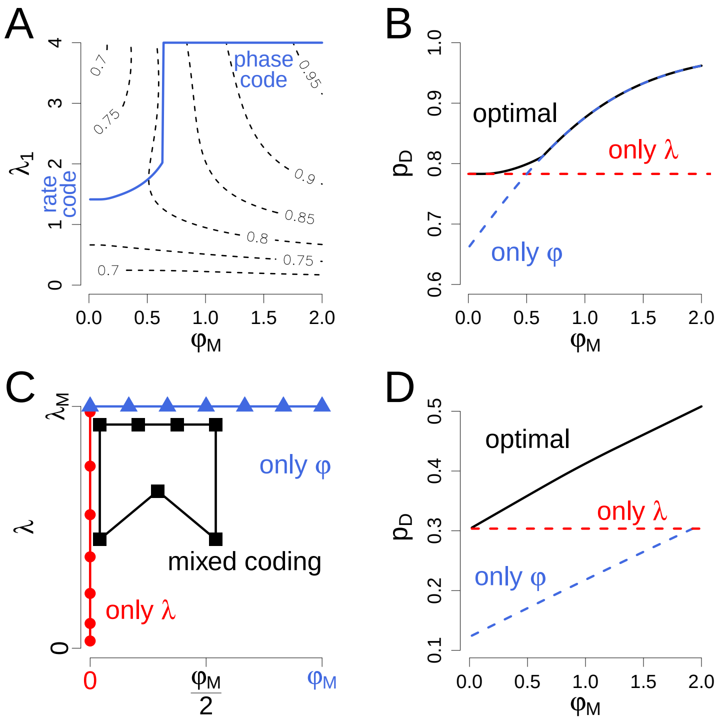


Figure 2.4: Optimal rate and phase parameters for $S = 2$ (A) and $S = 7$ (C) stimuli and corresponding detection probabilities (B,D). A. Value of λ_1 that maximizes p_D for $\lambda_M = 4$ in the case C_1, i.e., $\varphi_1 = 0$ and $\varphi_2 = \varphi_M$ as a function of φ_M. For $\varphi_M = 0$, we observe a pure rate code. For small φ_M, the optimal λ_1 remains constant as the phase is too small to contribute to stimulus detection. As φ_M increases, the optimal intermediate rate λ_1 increases and eventually takes the maximal rate. At this point, both stimuli are encoded with the same rate, resulting in a pure phase code. B. Maximal detection probability on basis of λ and φ (black) in comparison to a pure rate (red) and pure phase code (blue) for two stimuli. C. General optimal coding schemes for seven stimuli as a function of φ_M, indicated by colors. D. Maximal detection probability for seven stimuli for a pure rate (red) or pure phase (blue) code and a combination (black) of rate and phase.

Generally speaking, phase parameters can be advantageous for small and medium rates if the detection probability remains suboptimal for a pure rate code, which is particularly relevant for high numbers of stimuli. In addition, phase parameters can be even more advantageous in cases of high rates because in these cases, the mean spike time will be rather precise.

2.1.3 Similarity relations

In this section instead of maximizing the detection probability we want to minimize the weighted probabilities of false decision dependent on the distance of stimuli. In other words we want to maximize the probability making 'more or less' a correct decision. In Section 5 we observe in a setting of eight empirical neurons and 12 stimuli that these eight neurons confuse almost never stimuli, which are very different, cf. Figure 5.14 B. So besides a high detection probability it seems to be of extreme interest to minimize false decisions between stimuli, which are not close. Therefore, we introduce a linear error function, see Definition 2.1.9, and investigate how minimizing the linear error function changes the optimal parameters.

Again we assume a maximal rate λ_M and a maximal phase φ_M, i.e., $0 \leq \lambda_s \leq \lambda_M$ and $0 \leq \varphi_s \leq \varphi_M$ for all $s \in \{1, \ldots, S\}$. From now on stimuli are not ordered according to λ, but they are arranged uniformly on a circle, so that stimuli 1 and S have distance one, cf. Figure 5.14 A. Thus we assume a circular order with equal distances between stimuli, i.e, for S stimuli the distance between stimulus s_1 and s_2 is defined as

$$h_S(s_1, s_2) := \begin{cases} |s_1 - s_2| & \text{for } |s_1 - s_2| \leq \frac{S}{2}, \\ S - |s_1 - s_2| & \text{else.} \end{cases} \tag{2.16}$$

To keep notation compact we assume an even number S of stimuli.

Connection of misclassification and detection probability.
The probability of falsely detecting stimulus s_2 instead of the correct stimulus s_1 is denoted by $p_{s_1 s_2}$, i.e.,

$$p_{s_1 s_2} := \mathbb{P}_{s_1}(B \in A_{s_2}),$$

where $B = (N, \bar{X})$ and A_{s_2} is the acceptance region of stimulus s_2. Furthermore, let $p^{(\delta)}$ denote the average probability to misclassify two stimuli with a distance of δ, i.e.,

$$p^{(\delta)} := \frac{1}{S} \sum_{s,s' \colon h_S(s,s') = \delta} p_{ss'},$$

where $h_S(s, s')$ denotes the distance between stimuli s and s', cf. Equation 2.16. Considering a circular order of S stimuli with S even and equal distances between stimuli as shown in Figure 5.14 A, the overall detection probability p_D can also be obtained by a transformation using $p^{(\delta)}$, cf. Claim 2.1.8. Note that this connection only exists if either no nullstimulus exists or the nullstimulus is also part of the circular order.

Claim 2.1.8. *Assume an even number of stimuli and given rate parameters $\lambda_1, \ldots, \lambda_S$ and phase parameters $\varphi_1, \ldots, \varphi_S$ and that there is no additional nullstimulus. Then the detection* *probability p_D can be expressed as*

$$p_D = 1 - \sum_{\delta=1}^{S/2} p^{(\delta)}.$$

Proof. Let p_s denote the detection probability of stimulus s, i.e., the probability to decide for stimulus s if it is present. Then we obtain

$$p_D = \frac{1}{S}\sum_{s=1}^{S} p_s = \frac{1}{S}\left(S - \sum_{s\neq s'}^{S} p_{ss'}\right) = 1 - \sum_{\delta=1}^{S/2} p^{(\delta)}.$$

□

Linear detection error and its impact on optimal parameters.
Here we introduce the term *detection error* corresponding to the detection probability, which we want to minimize in this section. Thereby we naturally weight a misclassification dependent on the distance of the stimuli, but theoretically we have the problem of an additional nullstimulus. Here it is not obvious how to define the distance between stimulus s and the nullstimulus. Therefore, we weight a misclassification with the nullstimulus equally for every stimulus with maximal weight, cf. Definition 2.1.9.

Definition 2.1.9. *Given rate parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)$ and phase parameters $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_S)$ and an additional nullstimulus, the linear detection error e_D is defined as*

$$e_D\left(\boldsymbol{\lambda}, \boldsymbol{\varphi}\right) := \sum_{s_1,s_2=1}^{S} w_S(s_1,s_2)\cdot p_{s_1s_2} + \sum_{s=1}^{S} p_{s_10},$$

where $w_S(\cdot)$ is a linear weight function, i.e.,

$$w_S(s_1,s_2) := \frac{2}{S}\cdot h_S(s_1,s_2) \qquad \forall s_1,s_2 \in \{1,\ldots,S\}.$$

Remark 2.1.10. *The detection error e_D can be also written in terms of $p^{(\delta)}$, the average probability to misclassify two stimuli with a distance of δ, as*

$$\begin{aligned} e_D\left(\boldsymbol{\lambda}, \boldsymbol{\varphi}\right) &= \sum_{s_1,s_2=1}^{S} w_S(s_1,s_2)\cdot p_{s_1s_2} + \sum_{s=1}^{S} p_{s_10} \\ &= \sum_{s_1,s_2=1}^{S} \frac{2}{S}\cdot h_S(s_1,s_2)\cdot p_{s_1s_2} + \sum_{s=1}^{S} p_{s_10} \\ &= \sum_{\delta=1}^{S/2} 2\cdot\delta\cdot p^{(\delta)} + \sum_{s=1}^{S} p_{s_10}. \end{aligned}$$

Furthermore, note that we punish a false decision between stimuli with maximal distance of $S/2$ equally to a false decision for the nullstimulus, since $w_S(s_1,s_2) = 1$ if and only if $h_S(s_1,s_2) = S/2$. In case that the nullstimulus is present we never make a bad decision, as $p_0 = 1$.

Minimizing the linear detection error increases computational cost considerably, as every pairwise bad decision needs to be considered separately. Furthermore, for a decent number of stimuli, i.e., we are interested in 12 stimuli to obtain results comparable to the empirical neurons, the differences in detection probability or linear detection error of very different rate and phase parameter sets are quite small, whereby it is uncertain if an optimization algorithm will find the global maximum or minimum.
Therefore we first generate general insights how the optimal parameter set that minimizes the linear detection error will differ from the optimal parameter set that maximizes the detection probability. Second we test our results in case of 12 stimuli and rational choices of parameter sets according to Section 2.1.2.

In Figure 2.5 two parameter sets are illustrated, where the coordinates of the triangles represent the rate and phase parameter of a stimulus. The decision areas are drawn as black lines. Note that only integer spike numbers n can appear, but the grid points are linearly connected. As the number of spikes N and mean spike time $\bar{X}$ are sufficient for λ and φ, so far has only counted if the points (realizations) with the coordinates $(n, \bar{x})$ generated by stimulus s landed in the decision area of stimulus s or not. Now considering similarity relations it also counts, in which decision area the realization falls, if it misses the correct area.

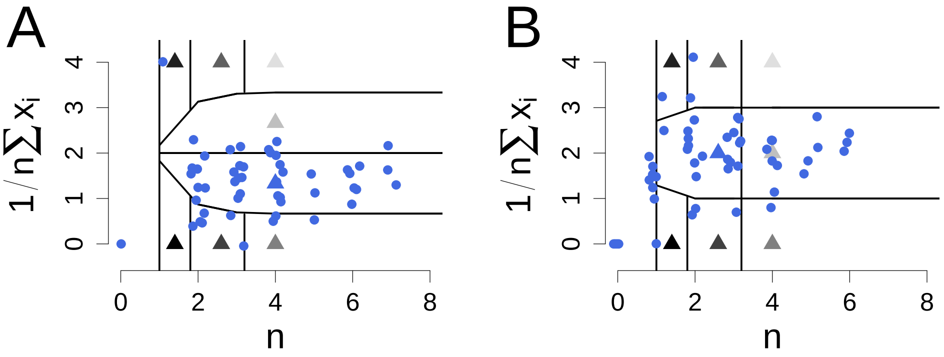


Figure 2.5: Acceptance regions for eight stimuli shown as black lines. The coordinates of the triangles represent the chosen rate and phase parameters in the λ-φ-plane. Each blue dot represents a realization according to the rate and phase parameters of the blue triangle. For better discrimination we use jitter for the observed spike numbers, as only discrete values occur. A realization is correctly assigned to the blue stimulus, if the blue dot is in the area of the blue triangle. A. Rectangular parameter structure, all stimuli are at the outer limit. B. Central parameter structure, one stimulus is in the mid of the other stimuli.

Basically to maximize the detection probability the parameters are chosen according to the following scheme: For small λ the deviation in direction of N is relatively small, but the deviation in direction of $\bar{X}$ is high. Equally for large λ the deviation in direction of N is relatively high, but the deviation in direction of $\bar{X}$ is small. So the points of the parameter set can be placed close to one another in direction of λ for small rates, and equally for high rates in direction of φ, cf. Figure 2.5.

Concerning the linear detection error the parameter set should be influenced in such a way, that bad decisions occur mostly to neighbored stimuli. So of special interest is, how minimizing the linear detection error effects the position of points, placed in the mid of many stimuli. This stimulus can not be obviously assigned to two neighbors. In Figure 2.5 A a parameter set of $S = 8$ stimuli is shown with no stimulus in the mid of the other stimuli. According to the rate and phase parameters of the blue triangle, random realizations (blue dots) are generated. Most of the dots, which are not in the acceptance region of the blue stimulus, fall into acceptance regions of direct neighbors. In Figure 2.5 B there is one stimulus with a middle rate parameter and a middle phase parameter (blue triangle). Misclassification of random realizations generated according to the middle stimulus is for the most part not restricted to direct neighbors, but occurs for almost all the other stimuli. Thus, as false decisions to close neighbors are less punished in case of the linear detection error, we would rather expect no mid stimulus in a parameter set that minimizes the linear detection error.

The illustration of Figure 2.5 can be confirmed in Figure 2.6. We consider $S = 12$ stimuli and

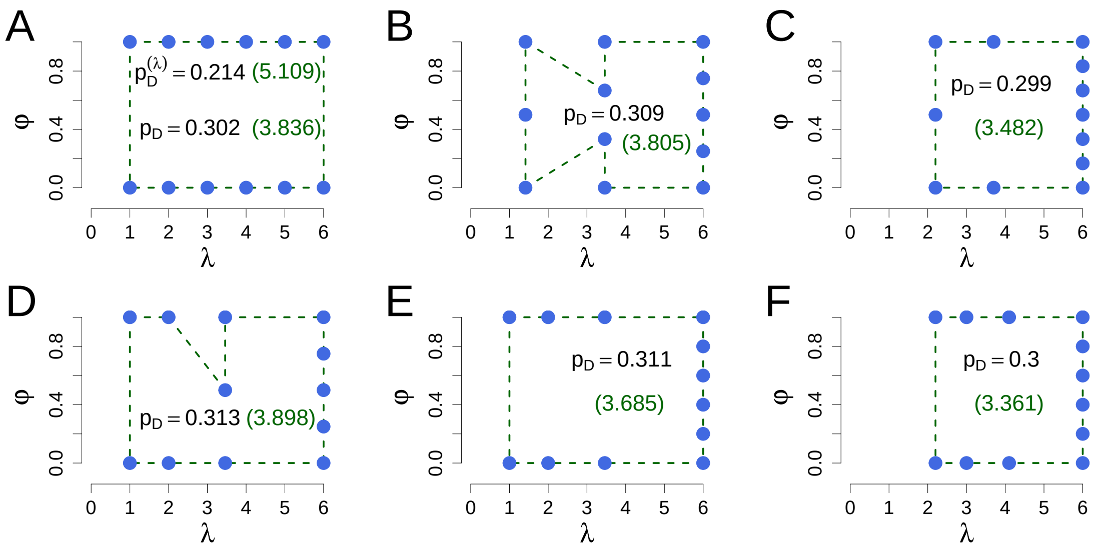


Figure 2.6: Detection probability and detection error for different parameter sets of $S = 12$ stimuli and $\lambda_M = 6$ and $\varphi_M = 1$. Each blue dot represents the rate and phase parameter of one stimulus. The green dotted line represents the circular order of the stimuli and connects adjacent stimuli. In black the detection probability is given, in green the detection error of the presented parameters. A. Optimal rate parameters in case of a pure rate code supported by the phase. The detection probability without the phase, i.e., $\varphi_M = 0$, is shown in the upper line indicated by $p_D^{(\lambda)}$. B (C). Parameter set that maximizes (minimizes) the detection probability (detection error) under the restriction of three different rate parameters. D (F). Parameter set that maximizes (minimizes) the detection probability (detection error) under the restriction of four different rate parameters. E. Slightly modified the parameter set of D, such that the stimulus with middle rate and phase is shifted to the maximal rate parameter and all six stimuli with λ_M are placed equidistant in φ from 0 to φ_M.

three basic different parameter sets. First in Figure 2.6 A we consider rate parameters, which maximize the detection probability in case of a pure rate code. As $\lambda_M = 6 \leq S$ we know according to Section 2.1.2.1 the optimal rate parameters are $1, \ldots, 6$ and we can distinguish only six stimuli. A pure rate code yields a detection probability of 0.214 and an detection error of 5.109. If we consider the same rate parameters, but use the phase ($\varphi_M = 1$) to distinguish respectively two stimuli in the same rate parameter, we can increase the detection probability by 41.1% to 0.302 and decrease the detection error by 24.9% to 3.836. In Figure 2.6 B we consider the parameter set that maximizes the detection probability under the restriction that only three different rate parameters are possible. Interestingly it is optimal to choose the rate parameters, which are optimal in case of pure rate code and three stimuli. Furthermore, every shift of a single rate or phase parameter decreases the detection probability. However, the increase in the detection probability compared to Figure 2.6 A is very small. Also the decrease in the detection error is quite small, but the shown parameter set should not be optimal in case of the detection error, as there exist two mid stimuli causing a higher error. If we place these stimuli at maximal rate we can decrease the detection error, cf. Figure 2.6 C. Shifting the not maximal rate parameters further to the maximal rate even decreases the detection error further, as this reduces the probability to falsely decide for the nullstimulus, which is punished with maximal weight. The loss in overall detection probability is negligible, i.e., 1%, but the decrease in the detection error is clearer, i.e., 10%
In Figure 2.6 D we consider the parameter set that maximizes the detection probability under the restriction that only four different rate parameters are possible. Interestingly, again it is optimal to choose the rate parameters, which are optimal in case of pure rate code and four stimuli. The increase in the detection probability compared to Figure 2.6 A is about 3.6% and there is no decrease in the detection error. There is one stimulus with middle rate and middle phase parameter. If we place this stimulus at maximal rate, the detection probability remains almost the same, but we can decrease the detection error by 5.5% to 3.685. If we also shift all parameters that are not maximal towards the maximal rate we can further decrease the detection error to 3.361 (8.8%), while the detection probability decreases only by 3.7%.
In summary maximizing the detection probability leads to stimuli with medium λ and medium φ, see Figure 2.6 B and D. Minimizing the linear error function leads to no stimulus with medium λ and medium φ and increases the not maximal rate parameters to avoid a misclassification with the nullstimulus, cf. 2.5 C and F. This effect is stable towards variations in the weight function, e.g. the square root of a linear weight function or similar weaker weighting with the same increasing weighting structure.

2.1.4 Linear discriminant analysis

In this section we give a short introduction to a well-known classification technique, the Linear Discriminant Analysis (LDA), apply it to our stimulus classification task and compare the result to the Bayesian decision rule. The primary purpose of LDA is to separate a sample of distinct groups by transforming the data to a different space that is optimal for distinguishing between the classes in the sense of the F-statistic.

We start with a short example how to distinguish between two multivariate normal distributions, if we use the Bayesian decision rule. Subsequently, we state the general assumptions of LDA and its optimization criteria. In Claim 2.1.14 we observe for $S = 2$ groups, that the solution of LDA equals the Bayesian decision rule in case of multivariate normal distributions, cf. Example 2.1.11. The general solution for $S \geq 2$ groups can be found in Lemma 2.1.16.

Afterwards we apply LDA to our stimulus classification task, where we observe the number of spikes and their spiketimes. Interestingly even if the assumption of equal covariance matrices is crucially violated, the decision bound for $S = 2$ stimuli using the Bayesian decision rule or LDA is almost the same (using the sum of spiketimes). However, introducing the nullstimulus shows that the classification with LDA is quite different from the Bayesian decision rule in some parameter cases. Increasing the number of stimuli seems to reduce the difference of both approaches.
However, LDA is no solution to easily find optimal parameters, since the optimization criterion is more difficult to handle and using only the first discriminant component loses a lot of information.

2.1.4.1 Introduction LDA

Example 2.1.11. *Consider two d-dimensional normally distributed random variables X_1 and X_2 with different means $\mu_1 \neq \mu_2$, but identical covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$, i.e.,*

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma) \quad \text{and} \quad X_2 \sim \mathcal{N}(\mu_2, \Sigma).$$

Again we think of X_1 as response to stimulus 1 and X_2 as response to stimulus 2. Equally, to Section 2.1.1 we assume that each stimulus is equally likely and our aim is to maximize the detection probability p_D. From Section 2.1.1, we already know that p_D is maximized by the Bayesian decision rule, i.e., we choose stimulus 1 if (with $f_s(\cdot)$ density of stimulus s)

$$f_1(x) > f_2(x).$$

Equating the normal densities with the assumption of equal covariance matrices yields

$$\begin{aligned} & \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) = \exp\left(-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right) \\ \Longleftrightarrow \quad & (x-\mu_1)^T\Sigma^{-1}(x-\mu_1) = (x-\mu_2)^T\Sigma^{-1}(x-\mu_2) \\ \Longleftrightarrow \quad & d(x) := x\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) = 0. \qquad (2.17) \end{aligned}$$

Thus the solution of all x, for which $f_1(x) = f_2(x)$, is a hyperplane.

Remark 2.1.12. *The hyperplane passes through the point $1/2(\mu_1 + \mu_2)$, since*

$$d\left(\frac{1}{2}(\mu_1+\mu_2)\right) = 0.$$

Remark 2.1.13. *Defining $a_{opt} := \Sigma^{-1}(\mu_1 - \mu_2)$, Equation 2.17 equals*

$$a_{opt}^T x = a_{opt}^T \frac{1}{2}(\mu_1+\mu_2).$$

The classical LDA by Fisher (Fisher, 1936) does not demand any distribution assumptions. LDA seeks to reduce dimensions, while preserving as much of the class discriminatory information as possible. However, if we desist from reducing dimensions, LDA provides in case of normal distributions the same result as using the Bayesian decision rule, see Example 2.1.11 and Claim 2.1.14.

LDA assumes a set of independent samples

$$\{X_{11},\dots,X_{u_1 1},X_{12},\dots,X_{u_2 2},\dots,X_{1S},\dots,X_{u_S S}\},$$

where u_s realizations belong to class s and it is known which class each realization belongs to. First we consider the case of only two groups:

Claim 2.1.14. *Let $x_{11},\dots,x_{u_1 1}$ be independent realizations of a random variable $X_1 \in \mathbb{R}^d$, $d \geq 1$, and independent of X_1 let $x_{12},\dots,x_{u_2 2}$ be independent realizations of a random variable $X_2 \in \mathbb{R}^d$, $d \geq 1$. We assume, the covariance matrix is identical for both groups, i.e. $\mathbb{C}ov[X_1] = \mathbb{C}ov[X_2]$. Furthermore, let*

$$\bar{x}_s := \frac{1}{u_s}\sum_{i=1}^{u_s} x_{is}, \qquad \textit{for } s = 1,2,$$

and define the between-class scatter matrix as

$$B = (\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)^T$$

and the within-class scatter matrix

$$\hat{\Sigma} = \hat{\Sigma}_1 + \hat{\Sigma}_2$$

with

$$\hat{\Sigma}_s = \sum_{i=1}^{u_s} (x_{is} - \bar{x}_s)(x_{is} - \bar{x}_s)^T, \qquad \textit{for } s = 1,2.$$

Then for $a \in \mathbb{R}^d$ the Fisher criterion

$$F(a) = \frac{a^T B a}{a^T \hat{\Sigma} a}$$

is maximized by

$$a_{opt} = \hat{\Sigma}^{-1}(\bar{x}_1 - \bar{x}_2).$$

Proof. The first derivative of $F(a)$ yields

$$\begin{aligned}
\frac{\partial}{\partial a}F(a) &= \frac{a^T\hat{\Sigma}a\frac{\partial a^T B a}{\partial a} - a^T B a\frac{\partial a^T\hat{\Sigma}a}{\partial a}}{\left(a^T\hat{\Sigma}a\right)^2} \\
&= \frac{a^T\hat{\Sigma}aa^T\left(B + B^T\right) - a^T Baa^T\left(\hat{\Sigma} + \hat{\Sigma}^T\right)}{\left(a^T\hat{\Sigma}a\right)^2} \\
&= \frac{a^T\hat{\Sigma}a2a^T B - a^T Ba2a^T\hat{\Sigma}}{\left(a^T\hat{\Sigma}a\right)^2},
\end{aligned}$$

since B and $\hat{\Sigma}$ are symmetric matrices. Equating to zero and dividing by $a^T\hat{\Sigma}a$ yields

$$\begin{aligned} & 2a^TB - \frac{a^TBa}{a^T\hat{\Sigma}a}2a^T\hat{\Sigma} = 0 \\ \Longrightarrow \quad & a^TB - F(a)a^T\hat{\Sigma} = 0 \\ \Longleftrightarrow \quad & Ba - F(a)\hat{\Sigma}a = 0 \\ \Longrightarrow \quad & \hat{\Sigma}^{-1}Ba - F(a)a = 0. \end{aligned}$$

Solving the generalized eigenvalue problem yields

$$\arg\max_a \frac{a^TBa}{a^T\hat{\Sigma}a} = \hat{\Sigma}^{-1}\left(\bar{x}_1 - \bar{x}_2\right).$$

□

Remark 2.1.15. *In LDA we maximize the difference between the projected means $a^T\bar{x}_1$ and $a^T\bar{x}_2$ normalized by a measure of the within-class scatter, since the difference between the projected means can be expressed as*

$$\left(a^T\bar{x}_1 - a^T\bar{x}_2\right)^2 = a^T\left(\bar{x}_1 - \bar{x}_2\right)\left(\bar{x}_1 - \bar{x}_2\right)^T a = a^TBa$$

and the scatter of the projections as

$$\sum_{i=1}^{u_s}\left(a^Tx_{is} - a^T\bar{x}_s\right)^2 = \sum_{i=1}^{u_s} a^T\left(x_{is} - \bar{x}_s\right)\left(x_{is} - \bar{x}_s\right)^T a = a^T\hat{\Sigma}_s a.$$

Let us now consider the general setting of $S > 2$ groups with total $U = \sum_{s=1}^{S} u_s$ observations. As in Claim 2.1.14 we define the between-class scatter matrix

$$B := \frac{1}{S}\sum_{s=1}^{S}\left(\bar{x}_s - \bar{x}\right)\left(\bar{x}_s - \bar{x}\right)^T,$$

where $\bar{x} := \frac{1}{U}\sum_{s=1}^{u_s} u_s\bar{x}_s$. The within-class scatter matrix generalizes as

$$\hat{\Sigma} := \frac{1}{U-S}\sum_{s=1}^{S}\hat{\Sigma}_s \quad \text{with} \quad \hat{\Sigma}_s = \sum_{i=1}^{u_s}\left(x_{is} - \bar{x}_s\right)\left(x_{is} - \bar{x}_s\right)^T.$$

Again the Fisher criterion is

$$F(a) = \frac{a^TBa}{a^T\hat{\Sigma}a}$$

and as the following Lemma tells us, is maximized by the eigenvector of $\hat{\Sigma}^{-1}B$ corresponding to the largest eigenvalue.

Lemma 2.1.16. *Let B and $\hat{\Sigma}$ be symmetric $d \times d$ matrices and $\hat{\Sigma}$ be positive semi definite. Then a^TBa is maximized by the eigenvector of $\hat{\Sigma}^{-1}B$ corresponding to the largest eigenvalue β_1 of $\hat{\Sigma}^{-1}B$ under the restriction $a^T\hat{\Sigma}a = 1$. This direction is called first discriminant component. Furthermore,* $\max_a a^TBa = \beta_1$.

Proof. See Mardia et al. (1979) □

Remark 2.1.17. *Further discriminant components are given by the eigenvectors of the next lower eigenvalues and maximize $F(a)$ under the restriction that the projections of the data are uncorrelated.*

Remark 2.1.18. *Maximum $m = \min\{S-1, d\}$ discriminant components can be determined, as B has maximum rank $S-1$ and $\hat{\Sigma}$ has maximum rank d.*

Remark 2.1.19. *Classification: Consider $r \leq m$ discriminant components $a_1, \dots, a_r$. Then a datum x is arranged in class c if*

$$\sum_{i=1}^{r} \left(a_i^T (x - \bar{x}_c)\right)^2 = \min_{s=1,\dots,S} \sum_{i=1}^{r} \left(a_i^T (x - \bar{x}_s)\right)^2 .$$

2.1.4.2 Application of LDA

Now our aim is to apply the LDA to our spiketrain setting where a stimulus specific Poisson distributed number of spikes and normally distributed spike times are used to distinguish between S stimuli and compare the detection probability resulting from LDA to the Bayesian decision rule.
In the LDA approach normally the matrices B and $\hat{\Sigma}$ are estimated. However, we know the real parameters λ_s and φ_s in our setting, so we do not need to estimate the mean or covariance matrices. To compare the outcome of LDA and Bayesian decision rule, it is also cleaner to use in both cases the known parameters, instead of estimating them only in the LDA scenario.

We recall, that we observe for stimulus s

$$Z_s = \begin{pmatrix} N_s \\ X_s \end{pmatrix},$$

where $N_s \sim Pois(\lambda_s)$ is the number of spikes and given $\{N_s = n_s\}$

$$X_s := \sum_{i=1}^{n_s} X_{is} \sim \mathcal{N}(n_s\varphi_s, n_s),$$

as $X_{is} \sim \mathcal{N}(\varphi_s, 1)$, is the sum of independent spike times. We will also consider $\bar{X}_s := \frac{1}{n_s} X_s$, where we define $\bar{X}_s := 0$ if $n_s = 0$. Notice that this choice does not change the decision criteria of the Bayesian decision rule and simplifies the application of LDA.

$S = 2$ **stimuli**
In the following we consider $S = 2$ stimuli (without the nullstimulus) and use Claim 2.1.14 to determine the decision rule. Therefore we first calculate the covariance matrix Σ_s separately for each group s (cf. Claim 2.1.20), which helps us to determine the pooled covariance matrix Σ, which is needed for the application of LDA, see Claim 2.1.22.

Claim 2.1.20. *Let $Z_s = (N_s, X_s)^T$ where $N_s \sim Pois(\lambda_s)$ and given $\{N_s = n_s\}$, we choose $X_s \sim \mathcal{N}(n_s\varphi_s, n_s)$. Then the covariance matrix of Z_s is given by*

$$\Sigma_s = \begin{pmatrix} \lambda_s & \lambda_s\varphi_s \\ \lambda_s\varphi_s & \lambda_s\left(1 + \varphi_s^2\right) \end{pmatrix}.$$

Proof. It is well-known that

$$\mathbb{V}ar[N_s] = \lambda_s$$

and the law of total variance yields

$$\begin{aligned}
\mathbb{V}ar[X_s] &= \mathbb{E}[\mathbb{V}ar[X_s \mid N_s]] + \mathbb{V}ar[\mathbb{E}[X_s \mid N_s]] \\
&= \mathbb{E}[N_s] + \mathbb{V}ar[\varphi_s N_s] \\
&= \lambda_s + \varphi_s^2 \lambda_s = \lambda_s \left(1 + \varphi_s^2\right).
\end{aligned}$$

As

$$\mathbb{E}[X_1] = \mathbb{E}[\mathbb{E}[X_1 \mid N_1]] = \mathbb{E}[\varphi_1 N_1] = \varphi_1 \lambda_1,$$

we again obtain with the law of total variance

$$\begin{aligned}
\mathbb{C}ov[N_s, X_s] &= \mathbb{E}[N_s X_s] - \mathbb{E}[N_s]\mathbb{E}[X_s] \\
&= \mathbb{E}[\mathbb{E}[N_s X_s \mid N_s]] - \lambda_s \varphi_s \lambda_s \\
&= \mathbb{E}[N_s \mathbb{E}[X \mid N_s]] - \lambda_s^2 \varphi_s \\
&= \mathbb{E}[N_s^2 \varphi_s] - \lambda_s^2 \varphi_s \\
&= \varphi_s \left(\mathbb{V}ar[N_s] + \mathbb{E}[N_s]^2 - \lambda_s^2\right) \\
&= \lambda_s \varphi_s.
\end{aligned}$$

□

Remark 2.1.21. *If we do not consider the sum of spike times, but the mean spike time, i.e., given $\{N_s = n_s\}$, we choose $\bar{X}_s \sim \mathcal{N}(\varphi_s, 1/n_s)$, the variables N_s and $\bar{X}_s$ are uncorrelated, as*

$$\mathbb{E}[N_s \bar{X}_s] = \mathbb{E}[\mathbb{E}[N_s \bar{X}_s \mid N_s]] = \mathbb{E}[N_s \mathbb{E}[\bar{X}_s \mid N_s]] = \mathbb{E}[N_s \varphi_s] = \lambda_s \varphi_s$$

and

$$\mathbb{E}[N_s]\mathbb{E}[\bar{X}_s] = \lambda_s \varphi_s.$$

But a problem concerning the mean spike time is, that we can not explicitly determine $\mathbb{V}ar[\bar{X}_s]$, which is needed for the covariance matrix, i.e.,

$$\begin{aligned}
\mathbb{V}ar[\bar{X}_s] &= \mathbb{E}[\mathbb{V}ar[\bar{X}_s \mid N_s]] + \mathbb{V}ar[\mathbb{E}[\bar{X}_s \mid N_s]] \\
&= \mathbb{E}\left[\frac{1}{N_s} \mathbb{1}_{\{N_s > 0\}} + 0 \cdot \mathbb{1}_{\{N_s = 0\}}\right] + \mathbb{V}ar[\varphi_s \mathbb{1}_{\{N_s > 0\}} + 0 \cdot \mathbb{1}_{\{N_s = 0\}}] \\
&= \sum_{i=1}^{\infty} \frac{1}{i} \frac{\lambda_s^i}{i!} \ell^{-\lambda_s} + \varphi_s^2 e^{-\lambda_s} \left(1 - e^{-\lambda_s}\right).
\end{aligned}$$

However, in the pooled version, the covariance of N and $\bar{X}$ is not zero, cf. Remark 2.1.23.

Claim 2.1.22. *Let $G \sim \mathcal{U}nif\{1, 2\}$ and given $\{G = s\}$, $s = 1, 2$, we choose*

$$Z = (N, X)^T,$$

where $N \sim Pois(\lambda_s)$ and given $\{N = n\}$, we choose $X \sim \mathcal{N}(n\varphi_s, n)$. Then the covariance matrix of Z is given by

$$\Sigma = \begin{pmatrix} \frac{1}{2}(\lambda_1 + \lambda_2) + \frac{1}{4}(\lambda_1 - \lambda_2)^2 & \mathcal{C}ov[X, N] \\ \mathcal{C}ov[X, N] & \frac{1}{2}\left(\lambda_1\left(1 + \varphi_1^2\right) + \lambda_2\left(1 + \varphi_2^2\right)\right) + \frac{1}{4}\left(\lambda_1\varphi_1 - \lambda_2\varphi_2\right)^2 \end{pmatrix},$$

where

$$\mathcal{C}ov[X, N] = \frac{1}{2}\left(\varphi_1\left(\lambda_1 + \lambda_1^2\right) + \varphi_2\left(\lambda_2 + \lambda_2^2\right)\right) - \frac{1}{4}(\lambda_1 + \lambda_2)(\varphi_1 + \lambda_1 + \varphi_2\lambda_2).$$

Proof. Using the results of Claim 2.1.20 for both stimuli yields

$$\begin{aligned} \mathbb{V}ar[N] &= \mathbb{E}[\mathbb{V}ar[N \mid G]] + \mathbb{V}ar[\mathbb{E}[N \mid G]] \\ &= \mathbb{E}\left[\mathbb{1}_{\{G=1\}} \cdot \lambda_1 + \mathbb{1}_{\{G=2\}} \cdot \lambda_2\right] + \frac{1}{2}\left(\lambda_1 - \frac{1}{2}(\lambda_1 + \lambda_2)\right)^2 + \frac{1}{2}\left(\lambda_2 - \frac{1}{2}(\lambda_1 + \lambda_2)\right)^2 \\ &= \frac{1}{2}\lambda_1 + \frac{1}{2}\lambda_2 + \frac{1}{2}\left(\frac{1}{2}\lambda_1 - \frac{1}{2}\lambda_2\right)^2 + \frac{1}{2}\left(\frac{1}{2}\lambda_2 - \frac{1}{2}\lambda_1\right)^2 \\ &= \frac{1}{2}\left(\lambda_1 + \lambda_2\right) + \frac{1}{4}\left(\lambda_1 - \lambda_2\right)^2 \end{aligned}$$

and

$$\begin{aligned} \mathbb{V}ar[X] &= \mathbb{E}[\mathbb{V}ar[X \mid G]] + \mathbb{V}ar[\mathbb{E}[X \mid G]] \\ &= \mathbb{E}\left[\mathbb{1}_{\{G=1\}} \cdot \lambda_1\left(1 + \varphi_1^2\right) + \mathbb{1}_{\{G=2\}} \cdot \lambda_2\left(1 + \varphi_2^2\right)\right] + \mathbb{V}ar[\mathbb{1}_{\{G=1\}} \cdot \lambda_1\varphi_1 + \mathbb{1}_{\{G=2\}} \cdot \lambda_2\varphi_2] \\ &= \frac{1}{2}\left(\lambda_1\left(1 + \varphi_1^2\right) + \lambda_2\left(1 + \varphi_2^2\right)\right) + \frac{1}{4}(\lambda_1\varphi_1 - \lambda_2\varphi_2)^2 \end{aligned}$$

and

$$\begin{aligned} \mathbb{C}ov[X, N] &= \mathbb{E}[NX] - \mathbb{E}[N] \\ &= \mathbb{E}[\mathbb{E}[NX \mid G]] - \frac{1}{2}(\lambda_1 + \lambda_2)\frac{1}{2}(\varphi_1\lambda_1 + \varphi_2\lambda_2) \\ &= \frac{1}{2}\varphi_1\left(\lambda_1 + \lambda_1^2\right) + \frac{1}{2}\varphi_2\left(\lambda_2 + \lambda_2^2\right) - \frac{1}{4}(\lambda_1 + \lambda_2)(\varphi_1\lambda_1 + \varphi_2\lambda_2). \end{aligned}$$

□

Remark 2.1.23. *As mentioned in Remark 2.1.21, if we consider the mean spike time, the covariance of N and X is zero treating each group separately. However, if we consider the setting of Claim 2.1.22 and determine the pooled covariance matrix, we obtain a covariance of N and X unequal zero, as*

$$\mathbb{E}[NX] = \mathbb{E}[\mathbb{E}[NX \mid G]] = \frac{1}{2}\lambda_1\varphi_1 + \frac{1}{2}\lambda_2\varphi_2$$

and

$$\mathbb{E}[N]\mathbb{E}[X] = \frac{1}{2}(\lambda_2 + \lambda_2)\frac{1}{2}(\varphi_1 + \varphi_2),$$

thus in general

$$\mathbb{C}ov[N, X] = \mathbb{E}[NX] - \mathbb{E}[N]\mathbb{E}[X] \neq 0.$$

Example 2.1.24. *Using Claim 2.1.14 and Claim 2.1.22 we are able to determine the LDA decision rule for the example parameter set $(\lambda_1, \varphi_1) = (2, 0)$ and $(\lambda_2, \varphi_2) = (4, 0.75)$. Claim 2.1.22 yields*

$$\Sigma = \begin{pmatrix} 4 & 3 \\ 3 & 6.375 \end{pmatrix}.$$

With Claim 2.1.14 we can determine

$$a_{opt} = \Sigma^{-1}\left(\begin{pmatrix} \lambda_1 \\ \lambda_1\varphi_1 \end{pmatrix} - \begin{pmatrix} \lambda_2 \\ \lambda_2\varphi_2 \end{pmatrix}\right) = \frac{1}{16.5}\begin{pmatrix} -3.75 \\ -6 \end{pmatrix}.$$

As $\mathbb{E}[Z] = \begin{pmatrix} 3 \\ 1.5 \end{pmatrix} =: \bar{z}$, we obtain for the projected mean

$$a_{opt}^T \bar{z} = -\frac{20.25}{16.5} \approx -1.23.$$

Overall this yields for the LDA decision line

$$x = \frac{a_{opt}^{(1)}}{a_{opt}^{(2)}} n + \frac{a_{opt}^T \bar{z}}{a_{opt}^{(2)}} = -0.625n + 3.375.$$

The comparison of the Bayesian decision rule with the LDA decision rule for $S = 2$ stimuli with parameter set of Example 2.1.24 can be found in Figure 2.7 B. The LDA decison rule is a straight line. Interestingly considering the sum of the spike times the Bayesian decision rule is also a straight line (red dashed) and is almost identical to the LDA decision rule. So in the setting of two stimuli (without the nullstimulus) we can use LDA instead of the Bayesian decision rule to maximize the detection probability, even if the assumption of equal covariance matrices is crucially violated. Only for a maximal phase larger than 1, the detection probability based on LDA is slightly lower than based on the Bayesian decision rule, cf. Figure 2.7 D.

Using Remark 2.1.21 we can analogously (numerically with R) to Example 2.1.24 calculate the LDA decision line, if we consider the mean spike time, see Figure 2.7 A. Here the Bayesian decision line looks quite different from the LDA decision line. However, both decision lines distinguish the two stimuli almost the same. Furthermore, the detection probability of applying LDA to the sum of spike times or the mean spike time results in almost the same detection probability.

An important intuition when applying LDA is, that we project the data to a subspace, which is optimal (in sense of the Fisher criterion) to distinguish between the stimuli. In case of $S = 2$ stimuli and the parameter set of Example 2.1.24 the projections are illustrated in Figure 2.7 C. If the projected data is to the left of the projected expected value $a_{opt}^T \bar{z}$ we decide for the green stimulus, otherwise we decide for the blue stimulus.

$S > 2$ **stimuli**

In the following we consider $S > 2$ stimuli and use Lemma 2.1.16 to determine the detection

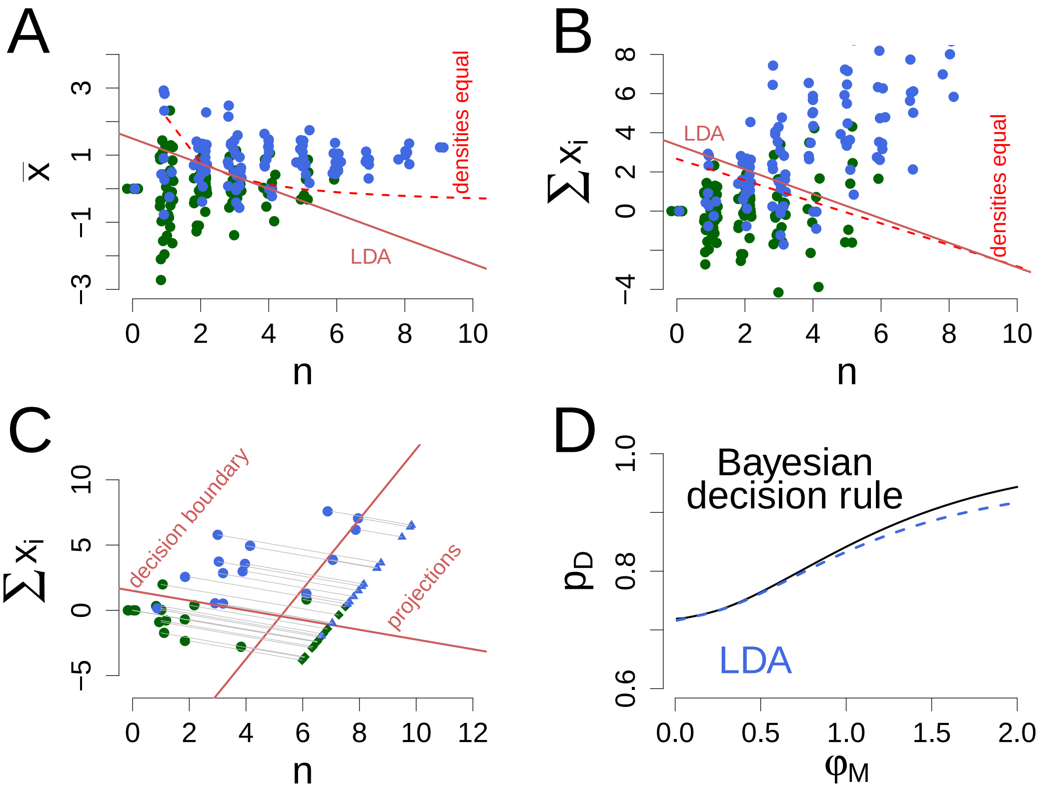


Figure 2.7: A and B. We consider the parameter set of Example 2.1.24 with $(\lambda_1, \varphi_1) = (2, 0)$ and $(\lambda_2, \varphi_2) = (4, 0.75)$ and apply LDA to distinguish between the two stimuli. The red dashed line represents the Bayesian decision rule, which gives the position of both densities equal. In dark red the decision line resulting of LDA is drawn. A. We consider the mean spiketime. B. We consider the sum of spike times. C. Visualization of the projections in LDA, cf. Claim 2.1.14. D. Detection probability resulting of the Bayesian decision rule compare to the detection probability resulting from the LDA.

probability resulting from LDA. Therefore, we recall the results of the case $S = 2$ stimuli and determine the required matrices

$$B = \frac{1}{S}\sum_{s=1}^{S}\left(\begin{pmatrix}\lambda_s\\ \lambda_s\varphi_s\end{pmatrix} - \bar{z}\right)\left(\begin{pmatrix}\lambda_s\\ \lambda_s\varphi_s\end{pmatrix} - \bar{z}\right)^T,$$

with

$$\bar{z} = \frac{1}{S}\sum_{s=1}^{S}\begin{pmatrix}\lambda_s\\ \lambda_s\varphi_s\end{pmatrix}$$

and according to Claim 2.1.20

$$\Sigma = \frac{1}{S}\sum_{s=1}^{S}\Sigma_s$$

with

$$\Sigma_s = \begin{pmatrix} \lambda_s & \lambda_s \varphi_s \\ \lambda_s \varphi_s & \lambda_s \left(1 + \varphi_s^2\right) \end{pmatrix}.$$

With Lemma 2.1.16 and Remark 2.1.17 we are able to determine the discriminant components. As $d = 2$ we have in case of $S > 2$ stimuli two discriminant components, cf. Remark 2.1.18. Using the classification rule of Remark 2.1.19 we can compare using both discriminant components with only the first component.

In Figure 2.8 we calculate the detection probability resulting from the Bayesian decision rule (black line) compared to resulting from the use of LDA dependent on the maximum phase. The detection probability based on both discriminant components is shown in blue, based on only the first component in green. We use the optimal parameter set, which we have determined in Section 2.1.2.3.

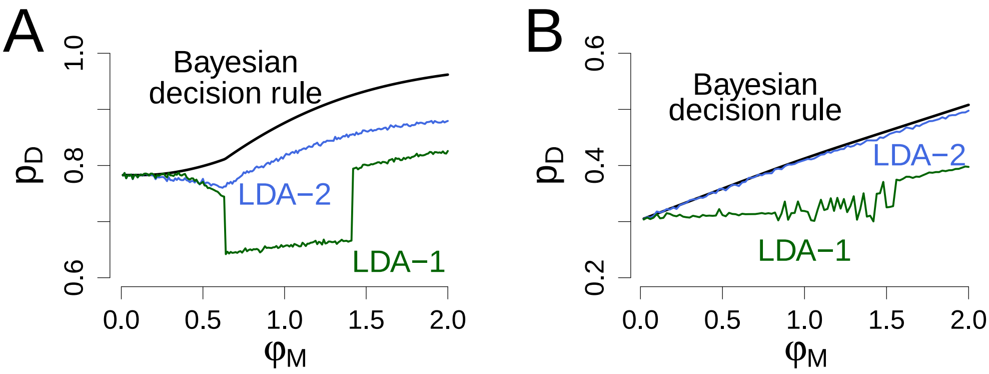


Figure 2.8: Detection probability resulting from the Bayesian decision rule (black line) compared to using classification by LDA dependent on the maximum phase. We use 1000 simulations per each phase value and choose the optimal parameter set. The detection probability based on both discriminant components ('LDA-2') is shown in blue, based on only the first component ('LDA-1') in green. A. $S = 2$ stimuli plus nullstimulus. B. $S = 7$ stimuli plus nullstimulus.

In Figure 2.8 A we consider $S = 2$ stimuli plus nullstimulus. For small phases up to $\varphi_M = 0.4$ LDA and Bayesian decision rule yield almost the same detection probability. In this range the first discriminant component is even enough, as the coding is mainly based on the rate. For larger phases the detection probability based on the Bayesian decision rule significantly differs from the detection probability based on LDA (two components). The difference increases if the maximum phase increases. Using only the first discriminant component we observe a significant drop in the detection probability at $\varphi_M \approx 0.6$. Here the coding structure changes from a rate and phase code to a pure phase code. As we use the rate to detect the nullstimulus, we still need both discriminant components. The first discriminant component detection rate recovers at $\varphi_M \approx 1.5$, but remains below the rate for LDA based on both components.
In Figure 2.8 A we consider $S = 7$ stimuli plus nullstimulus. Here Bayesian decision rule and LDA (both components) yields almost the same detection probability, differing slightly for

large phases. Already for small phases the second discriminant component carries important information. The difference of LDA using both components compared to only first component increases with increasing the maximum phase.
However, LDA is no solution to easily find optimal parameters, since the optimization criterion is more difficult to handle and using only the first discriminant component loses a lot of information.

2.1.5 Two Oscillation Cycles

For the previous results we have considered only one oscillation cycle in our GLO-Model (Figure 1.2 orange box) and asked for the optimal parameter set to maximize the detection probability or minimize the detection error. Now we want to explore in which manner two oscillation cycles in our GLO-Model affect the previous results. Again we assume we know the start time of each cycle (deterministic background beat), but now we have an additional uncertainty which cycle each spike belongs to. In case of one cycle we had to bound the maximal phase φ_M. Considering multiple oscillation cycles the maximal phase should adjust by itself for a given period μ_B, as otherwise it is hard to match the spikes to the correct oscillation cycle and the phase is of no use. If the phase is restricted and $\varphi_M \ll \mu_B$, we know for sure which oscillation cycle each spike belongs to. Thus to determine the rate and phase parameters that maximize the detection probability we can be draw from the results of Section 2.1.2, i.e., in case of two cycles and a maximal rate of λ_M it its equivalent to consider one cyle with maximal rate $2\lambda_M$, as the two independent inhomogeneous Poisson processes, with identical parameters, add to one inhomogeneous Poisson process with double maximal rate. So here it is of special interest that the maximal phase φ_M adjust itself to ensure a maximal detection probability.

It turns out that the detection probability is constant for μ_B/σ and $(\varphi_M - \varphi_{\min})/\sigma$ constant, see Figure 2.9 A and Lemma 2.1.26. Therefore, in this section we do not assume $\sigma = 1$, but consider the rate and phase parameters that maximize the detection probability dependent on $\sigma \in \mathbb{R}^+$, as it seems more natural to let μ_B fixed. Naturally the detection probability remains shift-invariant, so w.l.o.g $\varphi_{\min} = \min_s \varphi_s = 0$. For calculations, we assume a deterministic and known oscillation length μ_B and two oscillation cycles. Of particular interest is the influence of μ_B/σ on the maximal detection probability. Analogous to one oscillation cycle and one neuron the optimal phase parameters in a pure phase code are chosen equidistant, but here $\varphi_M = (S-1)/S\mu_B$, S the number of stimuli, as the distance to the following oscillation cycle need to be considered. So as the oscillation length μ_B is assumed to be fixed, the precision of the spike timing σ is the crucial parameter (in addition to the maximal rate).

Due to additional uncertainty which spike belongs to which oscillation cycle, the impact of the phase decreases compared to one oscillation cycle and one neuron with double maximal rate. Extending the model to stochastic oscillation would further decrease the impact of the phase. In addition, if the background oscillation is not observable the phase can only be used for at least two neurons. In practice, we do not know the starting point of each oscillation cycle, but with multiple neurons we are able to use the information of the phase difference of the spike times among the neurons.

Formally we consider two oscillation cycles in our GLO-model with equal deterministic length μ_B, i.e., $\sigma_B = 0$, cf. Definition 1.1.10 for one neuron. Thus w.l.o.g. the first cycle starts at time $t = 0$ and the second cycle at $t = \mu_B$. Analog to Section 2.1.1 we now consider two inhomogeneous Poisson processes, the first one with intensity

$$\rho_{1s}(t) = \frac{\lambda_s}{\sqrt{2\pi}} \exp\left(-\frac{(\varphi_s - t)^2}{2\sigma^2}\right), s \in \{1, \dots, S\},$$

and the second with intensity

$$\rho_{2s}(t) = \frac{\lambda_s}{\sqrt{2\pi}} \exp\left(-\frac{(\varphi_s - t - \mu_B)^2}{2\sigma^2}\right), s \in \{1, \dots, S\}.$$

Our decision task remains almost the same as in Section 2.1.1: Now we just observe two oscillation cycles, but we do not know which spike belongs to which cycle. Again we want to decide for the stimulus, which is most likely for the observed realization to maximize the detection probability $p_D = 1/S \sum_{s=1}^{S} \mathbb{P}_s(B \in A_s)$, as we still assume that all stimuli are equally likely. Now it is more complicated to determine the optimal acceptance regions $A_1, \dots, A_S$. Still we want to apply the Bayesian decision rule and assume for simplicity that the spikes before the first cycle and after the second cycle are still observable. In Claim 2.1.25 we state when we choose stimulus 1 according to the Bayesian decision rule for a given observation of n spikes and spike times $x_1, \dots, x_n$.

Claim 2.1.25. *Given S stimuli and rate parameters $\lambda_1, \dots, \lambda_S$ and phase parameters $\varphi_1, \dots, \varphi_S$ and assuming two oscillation cycles with deterministic length μ_B. According to the Bayesian decision rule for an observation of n spikes and spike times $x_1, \dots, x_n$ we decide for stimulus 1 if*

$$L_1^{\max} > L_s^{\max} \quad \forall s \in \{2, \dots, S\},$$

with

$$L_s^{\max} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \max_{j=0,\dots,n} \frac{\lambda_s^n}{j!(n-j)!} e^{-2\lambda_s} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{j}(x_i - \varphi_s)^2} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=j+1}^{n}(x_i - \mu_B - \varphi_s)^2}. \tag{2.18}$$

Proof. According to the Bayesian decision rule and equally likely stimuli we choose the stimulus, which is most likely for the observed spike sequence. Therefore, we need to decide for each spike which oscillation cycle it is belonging to, separately for each stimulus. So w.l.o.g. the maximum likelihood of stimulus 1 for a given realization $n, x_1, \dots, x_n$ is

$$L_1^{\max} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \max_{u \in \Omega} \frac{\lambda_1^{|u|}}{|u|!} e^{-\lambda_1} e^{-\frac{1}{2\sigma^2}\sum_{i \in u}(x_i - \varphi_1)^2} \cdot \frac{\lambda_1^{|u^c|}}{|u^c|!} e^{-\lambda_1} e^{-\frac{1}{2\sigma^2}\sum_{i \in u^c}(x_i - \mu_B - \varphi_1)^2},$$

where $\Omega := \wp(\{1, \dots, n\})$, with $\wp()$ the power set, and $u^c := \{i \in \{1, \dots, n\} : i \notin u\}$. However, the subset u yielding the maximal likelihood must be in the form of $\{1, \dots, j\}$, $j \leq N$, otherwise we could increase the likelihood by choosing $u' = \{1, \dots, |u|\}$. So Equation (2.18) holds for $L_s^{\max}$. □

For simplicity, we assumed that the spikes before the first oscillation and after the second oscillation are still observable. This assumption leads to a shift of the optimal maximal phase off the equidistant choice, as beyond the oscillation cycle is enough space and the allocation of every spike is reliable, cf. Figure 2.9 B. For this purpose we calculated the detection probability using Claim 2.1.25 for $S = 2$ stimuli and $\lambda_1 = \lambda_2$, $\sigma = 1$ and $\mu_B = 2$. Additionally we assumed a nullstimulus. As $\varphi_{\min} = 0$ the optimal phase parameters should be $\varphi_1 = 0$ and $\varphi_2 = \varphi_M/2 = 1$ or vice versa, if we would observe a infinite sequence of cycles. However, in case of two cycles and the assumption that outer spikes are still observable, the optimal φ_M within the cycle is close to 1.4. Actually an infinite value of φ_M would be optimal, but to get convincing results we set $\varphi_M := \mu_B/2$ in further calculations for $S = 2$ stimuli.

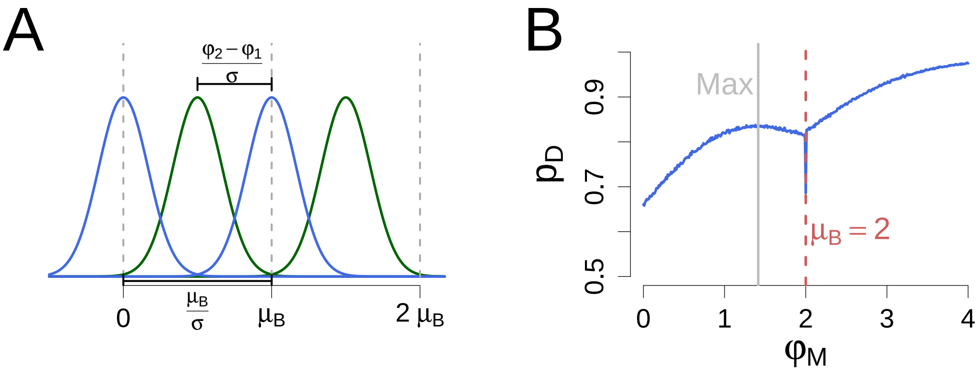


Figure 2.9: A. Considering $S = 2$ stimuli, spike time distribution of stimulus 1 in blue, of stimulus 2 in green, the detection probability is constant for μ_B/σ fixed (the relative space in each oscillation cycle) and $(\varphi_2 - \varphi_1)/\sigma$ fixed (the relative distance of the two stimuli distributions). For a calculation see Lemma 2.1.26. B. The optimal maximal phase φ_M for 2 stimuli and pure phase code for $\lambda = 2$, $\mu_B = 2$ and $\sigma = 1$ ($\varphi_{\min} = 0$). The optimal φ_M only depends on μ_B and should adjust itself at $\varphi_M = \mu_B/2$. The shift of the phase results of the assumption, that the spikes beyond the two considered oscillation cycle are still observable. However, to get convincing results we set $\varphi_M = \mu_B/2$ in further calculations.

The following lemma states that the detection probability is constant for μ_B/σ and $(\varphi_2-\varphi_1)/\sigma$ fixed as is illustrated in Figure 2.9 A.

Lemma 2.1.26. *Given S stimuli with rate parameters* $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_S)$ *and phase parameters* $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_S)$, *the detection probability is constant for* μ_B/σ *and* $(\varphi_M - \varphi_{\min})/\sigma$ *fixed, i.e.,*

$$p_D(\boldsymbol{\lambda}, \boldsymbol{\varphi}, \mu_B, \sigma) = p_D(\boldsymbol{\lambda}, c \cdot \boldsymbol{\varphi}, c \cdot \mu_B, c \cdot \sigma), \qquad \forall\, c \in \mathbb{R}.$$

Proof. Scaling $\boldsymbol{\varphi}$, μ_B and σ with a constant $\frac{1}{c}$ only leads to multiplying all likelihoods by c, i.e., with Claim 2.1.25 the likelihood of stimulus 1 is given by

$$\begin{aligned} L_1^{\max} &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \max_{j=0,\dots,n} \frac{\lambda_1^n}{j!(n-j)!} e^{-2\lambda_1} e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{j}(x_i-\varphi_1)^2+\sum_{i=j+1}^{n}(x_i-\mu_B-\varphi_1)^2\right)} \\ &= c\left(\frac{1}{\sqrt{2\pi(c\sigma)^2}}\right)^n \max_{j=0,\dots,n} \frac{\lambda_1^n}{j!(n-j)!} e^{-2\lambda_1} e^{-\frac{1}{2(c\sigma)^2}\left(\sum_{i=1}^{j}(cx_i-c\varphi_1)^2+\sum_{i=j+1}^{n}(cx_i-c\mu_B-c\varphi_1)^2\right)}. \end{aligned}$$

As we detect the stimulus with maximal likelihood, the scaling does not effect the detection probability. □

In Figure 2.10 we determine the parameters that maximize the detection probability for one neuron and two stimuli for two oscillation cycles by simulations. Since the optimal phase parameters are fixed for constant μ_B and naturally one stimulus is chosen with maximal rate, we only have to determine the lower rate $\lambda_{\min}$ dependent on the spike timing precision σ, i.e., w.l.o.g. $\lambda_1 = \lambda_{\min}$, $\lambda_2 = \lambda_M$, $\varphi_1 = 0$ and $\varphi_2 = \varphi_M = \mu_B/2$. The optimal rate parameters in case of a pure rate code can be obtained easily based on the results of one oscillation cycle, see Remark 2.1.27.

Remark 2.1.27. *Assume a pure rate code, i.e.,* $\varphi_1 = \cdots = \varphi_S$*, and a given maximal rate* λ_M*. Furthermore let* $\lambda_1 \leq \cdots \leq \lambda_S = \lambda_M$ *denote the rate parameters that maximize the detection probability in case of two cycles and* $\tilde{\lambda}_1 \leq \cdots \leq \tilde{\lambda}_S = 2\lambda_M$ *denote the rate parameters that maximize the detection probability in case of one cycle. Then it holds that*

$$p_D(\lambda_1, \ldots, \lambda_S) = p_D(\tilde{\lambda}_1/2, \ldots, \tilde{\lambda}_S/2) \quad \textit{considering two cycles.}$$

In case of two cycles we observe in each cycle an independent Pois(λ_s)*-distributed random number, where the optimal allocation of the spikes to the cycle only depends on n and not on* λ_s *and is thus equal for all stimuli, cf. Equation* (2.18)*. So we can sum both Poisson numbers and obtain one Pois*$(2\lambda_s)$*-distributed random number and draw on the optimal solution in case of one cycle and a maximal rate of* $2\lambda_M$*.*

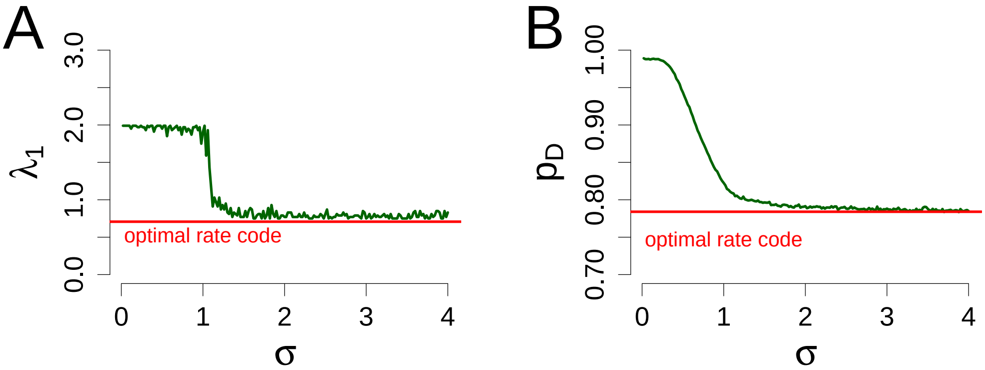


Figure 2.10: A. The optimal minimal rate parameter λ_1 (green line) dependent on the spike timing precision σ for 2 stimuli and two oscillation cycles and $\lambda_M = 2$, $\mu_B = 2$ and $\varphi_M = \mu/2$. In case of a small σ, i.e., the spikes are placed with a high precision, it is optimal to choose $\lambda_1 = \lambda_M$ and exploit the strong separation in the phase. Increasing σ results in a degradation of the discrimination in the phase and the minimal rate decreases for $\sigma \in [1, 1.2]$ very fast to the minimal rate of $\lambda_1 = \sqrt{2}/2$, already known from a optimal pure rate code in one oscillation cycle ($\boldsymbol{\lambda} = (\sqrt{2}, 4)$, cf. Remark 2.1.27). B. Maximal detection probability dependent on the spike timing precision σ for the optimal λ_1 (green line). In red the detection probability which results from an optimal pure rate code.

Figure A shows the impact of σ on the choice of the optimal minimal rate λ_1. The result is almost the same compared to one oscillation cycle. If σ is small, which corresponds to highly precise phases, we start with a pure phase code, i.e., $\lambda_1 = \lambda_M = 2$. Increasing σ, which is equivalent to decreasing the precision of φ_M, results for $\sigma = 1$ virtually in a jump in a minimal rate of $\lambda_1 = \sqrt{2}/2$, which maximizes the detection probability in case of a pure rate code, cf. Remark 2.1.27 and Section 2.1.2.1. Interestingly at a rate of $\lambda_1 = 1$, which is the optimal rate parameter in case of a pure rate code ($\lambda_M = 2$) and one oscillation cycle, the jump in the rate is interrupted and the rate decreases more slowly to the optimal minimal rate in case of a pure rate code (and two cycles).

Figure B illustrates the influence of the precision of the spike timing σ on the maximal detection probability p_D. We start with a detection probability of nearly one, as stimulus 1 and 2 can be distinguished almost always with highly precise phases. Only if there is no spike in both cycles for stimulus 1 or 2, we obtain a misclassification with the nullstimulus, i.e., $p_D = 1/3(1 + 2 \cdot (1 - (e^{-2})^2) \approx 0.988$. Increasing σ decreases the detection probability up to the detection probability of an optimal rate code.

Since $\varphi_M = \mu_B/2 = 1$, we need a sigma of 4/3 to compare the results to Section 2.1.2.3, where we analyzed one cycle and were interested in phases up to $\varphi_M = 0.75$ ($\sigma = 1$). However with a σ of 4/3 the optimal rate and phase code has almost the same rate parameters that are optimal for a pure rate code. In case of two cycle the optimal rate and phase code ($\lambda_1 \approx 0.89, \lambda_2 = 2, \varphi_1 = 0, \varphi_2 = 1, \sigma = 4/3$ yielding $p_D \approx 0.796$) increases the detection probability only by 1.7% compared to a pure rate code ($\lambda_1 = \sqrt{2}/2, \lambda_2 = 2, \varphi_1 = 0, \varphi_2 = 0, \sigma = 4/3$ yielding $p_D \approx 0.783$). In one cycle the optimal rate and phase code ($\lambda_1 = 4, \lambda_2 = 4, \varphi_1 = 0, \varphi_2 = 0.75$ yielding $p_D \approx 0.834$) increases the detection probability by 6.51% compared to a pure rate code ($\lambda_1 = \sqrt{2}, \lambda_2 = 4, \varphi_1 = 0, \varphi_2 = 0$ yielding $p_D \approx 0.783$).

In summary already for two deterministic oscillation cycles the ability of phase to increase the detection probability decreases compared to one cycle. Due to the additional uncertainty of the spike allocation to the correct oscillation cycle, the precision of the phase needs to be higher to increase the detection probability comparable to one cycle. However, the basic coding properties remain: Starting with a pure phase code in case of highly precise phases, the optimal minimal rate parameter decreases with decreasing precision of the phase up to the optimal pure rate code.

2.2 Two neurons

Here we investigate the probability of correct stimulus detection, p_D, for two neurons within a single oscillation cycle, as a function of the spiking parameters rate λ and phase φ. To this end we restrict to one cycle of our GLO-Model (Figure 1.2 extended orange box, green and blue firing intensity) and again assume we know the start time of the cycle. An illustration of the decision task in case of two stimuli can be found in Figure 2.11 A: For each neuron we observe the number and timing of spikes (red (dashed) bars) and need to decide, which stimulus caused the spiking output. Therefore, we assume we know the spiking intensity of each neuron and each stimulus. The spiking intensity of each stimulus and neuron directly corresponds to a rate and a phase parameter, which is shown in Figure 2.11 B. So our main objective is to analyze which rate and phase parameters maximize the detection probability p_D. Already for two neurons and two stimuli we have four rate and four phase parameters, which results in an expensive optimization especially for a high number of stimuli. However, we want to get a basic idea, if and for which numbers of stimuli and neurons imprecise phases can considerably increase the detection probability.

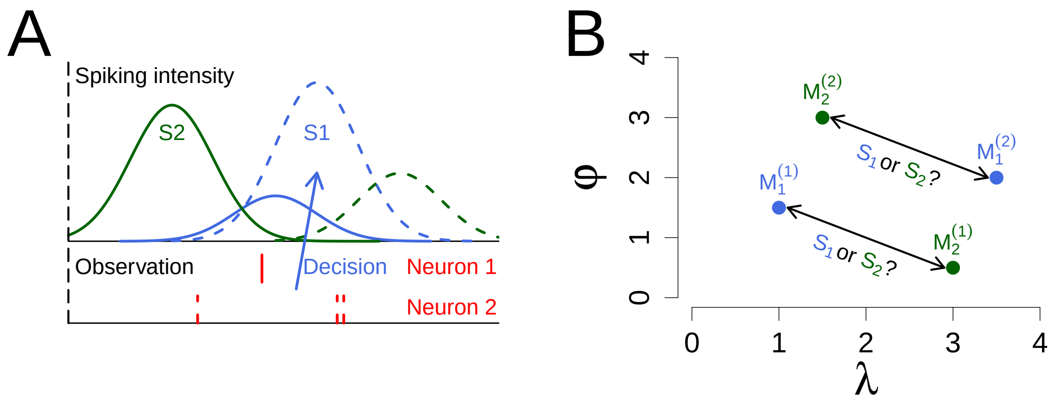


Figure 2.11: Decision Task for two stimuli and two neurons. A. For each neuron we observe the number and timing of spikes, the spikes of neuron 1 are illustrated as red bars and of neuron 2 as red dashed bars. The spiking intensity corresponding to stimulus 1 is shown in blue, to stimulus 2 in green. The spiking intensity belonging to neuron 2 is shown in dashed lines. In case of the shown observation we would decide for stimulus 1 (blue). B. According to our spiking model, cf. Equation (2.19), the spiking intensity results from a rate and a phase parameter, which are illustrated as four dots in the λ-φ-plane. Thereby $M_s^{(m)}$ denotes the rate and phase parameters of neuron m responding to stimulus s.

Formally, we consider a set $\{1, \ldots, S\}$ of $S \in \mathbb{N}$ stimuli and for each neuron $m \in \{1, \ldots, M\}$ of $M \in \mathbb{N}$ rate parameters $\boldsymbol{\lambda}^{(m)} = \left(\lambda_1^{(m)}, \ldots, \lambda_S^{(m)}\right)$, with $\lambda_s^{(m)} \geq 0 \, \forall s, m$ and phase parameters $\boldsymbol{\varphi}^{(m)} = \left(\varphi_1^{(m)}, \ldots, \varphi_S^{(m)}\right)$, with $\varphi_s^{(m)} \in \mathbb{R} \, \forall s, m$. We assume that the spiking response of each neuron $m \in \{1, \ldots, M\}$ within an oscillation cycle can be described by an inhomogeneous Poisson process with intensity

$$\rho_s^{(m)}(t) = \frac{\lambda_s^{(m)}}{\sqrt{2\pi}} \exp\left(-\frac{(\varphi_s^{(m)} - t)^2}{2}\right), s \in \{1, \ldots, S\}. \tag{2.19}$$

Thus we assume each neuron responses to stimulus s with an independent $Pois(\lambda_s^{(m)})$-distributed number $N_s^{(m)}$ of spikes, where the spike times $X_{is}^{(m)}$, $i = 1, \ldots, N_s^{(m)}$ are independent and $\mathcal{N}(\varphi_s^{(m)}, 1)$-distributed. Again the imprecision of spikes, σ^2, is set to 1 because only the relation of φ and σ is relevant, assuming that σ^2 is equal for all stimuli and neurons.

For $S = 2$ stimuli the acceptance regions are described in Claim 2.2.1. In the general case of $S > 2$ stimuli a numeric calculation has a high computational cost, cf. Remark 2.2.4, thus we determine the detection probability with Lemma 2.2.2 by simulations. In case of a pure rate code, i.e., all phases are equal, Corollary 2.2.3 states how to calculate the detection probability numerically. In the special case of $S = 2$ stimuli plus a nullstimulus, such a stimulus always exists in an optimal parameter set, cf. Lemma 2.2.5, we can use Claim 2.2.1 to calculate the detection probability numerically efficient, see Paragraph 'Special case of $M = 2$ neurons and $S = 2$ stimuli' in Section 2.2.1.

In Section 2.2.2 we explore the optimal rate and phase parameters in case of two neurons, starting with the insight that an optimal parameter set always contains a nullstimulus, see Section 2.2.2.1. With the numerically calculation for two stimuli (plus nullstimulus) we determine the optimal rate and phase parameters in Section 2.2.2.2. In Section 2.2.2.3 we use simulations to analyze, if small phases can increase the detection probability in case of three and four stimuli.

2.2.1 The detection probability

In order to derive the probability to detect the correct stimulus among S stimuli, we extend the notation of Section 2.1.1 to M neurons. A realization of the random vector $B_s^{(m)} = \left(N_s^{(m)}, \bar{X}_s^{(m)}\right)$, where $N_s^{(m)}$ denotes the number of spikes and $\bar{X}_s^{(m)} := \frac{1}{n^{(m)}} \sum_{i=1}^{n^{(m)}} X_{is}^{(m)}$ the mean spike time of neuron m, is denoted by

$$b^{(m)} := \left(n^{(m)}, \bar{x}^{(m)}\right) \in \mathbb{N} \times \mathbb{R}.$$

As $\bar{X}_s \sim \mathcal{N}(\varphi_s^{(m)}, \sigma^2/n^{(m)})$ given $\{N_s^{(m)} = n^{(m)}\}$, we define for a realization $b^{(m)}$

$$P_s\left(b^{(m)}\right) := \mathbb{P}\left(N_s^{(m)} = n^{(m)}\right) \phi_{\varphi_s^{(m)}, \sigma^2/n^{(m)}}\left(\bar{x}^{(m)}\right),$$

where $\phi_{\varphi_s^{(m)}, \sigma^2/n^{(m)}}$ denotes the density of the normal distribution with mean $\varphi_s^{(m)}$ and variance $\sigma^2/n^{(m)}$ at its argument.

For each neuron $m \in \{1, \ldots, M\}$ we have an observation space $\mathbb{N} \times \mathbb{R}$. Now the decision which stimulus is present depends on $M > 1$ neurons. Thus, the whole observation space is $(\mathbb{N} \times \mathbb{R})^M$, which we divide into S acceptance regions $A_1, \ldots, A_S$ such that we maximize the detection probability, i.e.,

$$p_D := \frac{1}{S} \sum_{s=1}^{S} \mathbb{P}_s(\mathbf{B} \in A_s),$$

where $\mathbb{P}_s(\mathbf{B} \in A_s) := \mathbb{P}(\mathbf{B}_s \in A_s)$ and $\mathbf{B}_s := \left(B_s^{(1)}, \ldots, B_s^{(M)}\right)$ for $s = 1, \ldots, S$.

Again according to the Bayesian decision rule (e.g., Camastra and Vinciarelli, 2015), the optimal set of acceptance regions assigns an observation $\mathbf{b} = \left(b^{(1)}, \ldots, b^{(M)}\right)$ to stimulus s if

$$\prod_{m=1}^{M} P_s^{(m)}\left(b^{(m)}\right) > \prod_{m=1}^{M} P_{s'}^{(m)}\left(b^{(m)}\right) \quad \forall s' \neq s, \tag{2.20}$$

or in short $P_s(\mathbf{b}) > P_{s'}(\mathbf{b})$ $\forall s' \neq s$, under the assumption that all stimuli are equally likely.

If all neurons have the same phase parameters for more than one stimulus, i.e., $\exists\, s \neq s' \in \{1, \ldots, S\}$ with $\varphi_s^{(1)} = \cdots = \varphi_s^{(M)} = \varphi_{s'}^{(1)} = \cdots = \varphi_{s'}^{(M)}$, it is possible due to the discreteness of the Poisson distribution that for an observation $\mathbf{b}$ multiple stimuli yield the same maximal $P_s(\mathbf{b})$, i.e., $\exists\, \tilde{S} \subset \{1, \ldots, S\}$ with $|\tilde{S}| \geq 2$ and $P_{\tilde{s}}(\mathbf{b}) = P_{\tilde{s}'}(\mathbf{b})$ $\forall\, \tilde{s}, \tilde{s}' \in \tilde{S}$ and $P_{\tilde{s}}(\mathbf{b}) > P_s(\mathbf{b})$ $\forall \tilde{s} \in \tilde{S}, s \notin \tilde{S}$. In this cases we assign this observation to the stimulus with the smaller sum of rate parameters, as assigning b to any of the stimuli $\tilde{s} \in \tilde{S}$ maximizes the detection probability.

Claim 2.2.1. *Let $\boldsymbol{\lambda}^{(1)}, \dots, \boldsymbol{\lambda}^{(M)}$, $\lambda_s^{(m)} \geq 0$ for all $m \in \{1, \dots, M\}$, $s \in \{1, \dots, M\}$ and $\boldsymbol{\varphi}^{(1)}, \dots, \boldsymbol{\varphi}^{(M)}$, $\varphi_s^{(m)} \in \mathbb{R}$ for all $m \in \{1, \dots, M\}$, $s \in \{1, \dots, M\}$ be rate and phase parameters for $S = 2$ stimuli and M neurons and let $\left|\left\{\varphi_1^{(1)}, \dots, \varphi_S^{(1)}, \dots, \varphi_1^{(M)}, \dots, \varphi_S^{(M)})\right\}\right| > 1$. Let $N^{(m)} = n^{(m)}$ be the number of spikes and $\bar{X}^{(m)} = \bar{x}^{(m)}$ be the mean observed spike time of neuron $m = 1, \dots, M$. W.l.o.g. the acceptance region of stimulus 1 is given by the set*

$$\begin{aligned} A_1 := \Bigg\{ \left((n^{(1)}, \bar{x}^{(1)}), \dots, (n^{(M)}, \bar{x}^{(M)})\right) \Bigg| \sum_{m=1}^{M} \Bigg[n^{(m)} \log \frac{\lambda_1^{(m)}}{\lambda_2^{(m)}} \\ - \frac{\sqrt{n^{(m)}}}{\sigma}(\varphi_2^{(m)} - \varphi_1^{(m)}) \left(\frac{\bar{x}^{(m)} - \varphi_1^{(m)}}{\sigma/\sqrt{n^{(m)}}} + \frac{\sqrt{n^{(m)}}}{\sigma} \frac{\varphi_1^{(m)} - \varphi_2^{(m)}}{2} \right) \Bigg] > \sum_{m=1}^{M} \lambda_1^{(m)} - \lambda_2^{(m)} \Bigg\}. \end{aligned} \tag{2.21}$$

Proof. For two stimuli, acceptance region A_1 is defined by the set of all b such that $P_1(\mathbf{b}) > P_2(\mathbf{b})$, or the set of all $(n^{(1)}, \bar{x}^{(1)}), \dots, (n^{(M)}, \bar{x}^{(M)})$ with

$$\prod_{m=1}^{M} \left(\frac{\lambda_1^{(m)}}{\lambda_2^{(m)}} \right)^{n^{(m)}} e^{\lambda_2^{(m)} - \lambda_1^{(m)}} e^{-\frac{n^{(m)}}{2\sigma^2}\left(\left(\bar{x}^{(m)} - \varphi_1^{(m)}\right)^2 - \left(\bar{x}^{(m)} - \varphi_2^{(m)}\right)^2\right)} > 1.$$

Applying the natural logarithm yields

$$\begin{aligned} & \sum_{m=1}^{M} n \log \frac{\lambda_1^{(m)}}{\lambda_2^{(m)}} - \frac{n^{(m)}(\varphi_2^{(m)} - \varphi_1^{(m)})}{\sigma^2} \left(\bar{x}^{(m)} - \frac{\varphi_1^{(m)} + \varphi_2^{(m)}}{2} \right) > \lambda_1^{(m)} - \lambda_2^{(m)} \\ \iff \quad & \sum_{m=1}^{M} n \log \frac{\lambda_1^{(m)}}{\lambda_2^{(m)}} - \frac{\sqrt{n^{(m)}}}{\sigma}(\varphi_2^{(m)} - \varphi_1^{(m)}) \left(\frac{\bar{x}^{(m)} - \varphi_1^{(m)}}{\sigma/\sqrt{n^{(m)}}} + \frac{\sqrt{n^{(m)}}}{\sigma} \frac{\varphi_1^{(m)} - \varphi_2^{(m)}}{2} \right) > \lambda_1^{(m)} - \lambda_2^{(m)}. \end{aligned}$$

□

General case of M neurons and S stimuli

With Claim 2.2.1 we basically have the information to calculate the detection probability: For $S \geq 2$ stimuli we need to consider the acceptance regions of all S stimuli simultaneously, cf. Equation (2.21). The calculation formula of the detection probability is given in Lemma 2.2.2.

Lemma 2.2.2. *Given S stimuli and M neurons with rate parameters $\boldsymbol{\lambda}^{(1)}, \dots, \boldsymbol{\lambda}^{(M)}$, $\lambda_s^{(m)} \geq 0$ for all $m \in \{1, \dots, M\}$, $s \in \{1, \dots, S\}$ and phase parameters $\boldsymbol{\varphi}^{(1)}, \dots, \boldsymbol{\varphi}^{(M)}$, $\varphi_s^{(m)} \in \mathbb{R}$ for all $m \in \{1, \dots, M\}$, $s \in \{1, \dots, S\}$. Let $p_s := \mathbb{P}_s(\mathbf{B} \in A_s)$ denote the probability to detect stimulus s correctly, thus the detection probability is $p_D = 1/S(p_1 + \dots + p_S)$. Then p_s is given by*

$$p_s = \mathbb{P}_s \left(G_{sr} > 0 \ \forall r \neq s, r \in \{1, \dots, S\} \right),$$

where

$$\begin{aligned} & G_{sr} := \\ & \sum_{m=1}^{M} \left[N^{(m)} \log \frac{\lambda_s^{(m)}}{\lambda_r^{(m)}} - \frac{\sqrt{N^{(m)}}}{\sigma}(\varphi_r^{(m)} - \varphi_s^{(m)}) \left(Z^{(m)} + \frac{\sqrt{N^{(m)}}}{\sigma} \frac{\varphi_s^{(m)} - \varphi_r^{(m)}}{2} \right) - \lambda_s^{(m)} + \lambda_r^{(m)} \right], \end{aligned}$$

with $Z^{(m)} \sim \mathcal{N}(0, 1)$ and $N^{(m)} \sim Pois\left(\lambda_s^{(m)}\right)$ for all $m \in \{1, \dots, M\}$ and independent.

Proof. We recall that the probability to detect stimulus s if it is present (p_s) is the same as the probability that Equation (2.21) from Claim 2.2.1 is met for all $r \neq s$. Furthermore we recognize that $\frac{\bar{x}^{(m)}-\varphi_1^{(m)}}{\sigma/\sqrt{n^{(m)}}} \sim \mathcal{N}(0,1)$ and that $\bar{X}^{(1)},\dots,\bar{X}^{(M)}$ and $\bar{N}^{(1)},\dots,\bar{N}^{(M)}$ are all independent. □

For an arbitrary number of stimuli $S > 2$ and not all phase parameters equal we determine the detection probability by simulation, as we can not calculate the detection probability numerically efficient, cf. Remark 2.2.4. In the special case of a pure rate code, i.e., $\varphi_s^{(m)} = 0\ \forall\, s, m$, the detection probability can be computed numerically efficient for any number of stimuli, cf. the following corollary.

Corollary 2.2.3. *Given S stimuli and M neurons with rate parameters $\boldsymbol{\lambda}^{(1)},\dots,\boldsymbol{\lambda}^{(M)}$, $\lambda_s^{(m)} \geq 0$ for all $m \in \{1,\dots,M\}$, $s \in \{1,\dots,M\}$ and let all phase parameters equal, i.e., $\left|\left\{\varphi_1^{(1)},\dots,\varphi_S^{(1)},\dots,\varphi_1^{(M)},\dots,\varphi_S^{(M)})\right\}\right| = 1$. Then the detection probability is given by*

$$p_D = \frac{1}{S} \sum_{n^{(1)},\dots,n^{(M)}=0}^{\infty} \frac{1}{n^{(1)}!\cdots n^{(M)}!} \max_{s\in\{1,\dots,S\}} \left(\prod_{m=1}^{M} \left(\lambda_s^{(m)}\right)^{n^{(m)}} e^{-\lambda_s^{(m)}} \right).$$

Proof. For a given realization $\mathbf{B}$ we choose the stimulus which is most likely for this event. As in case of a pure rate code only the spike numbers count, we need to sum up the maximal stimulus specific weights for all spike number combinations. □

Remark 2.2.4. *In case of $S > 2$ stimuli and $M > 1$ neurons we cannot determine the detection probability numerically efficient and need to use simulations: Analogous to the case of one neuron we need to determine the optimal decision areas by calculating a maximum or minimum over all stimuli, cf. Equation* (2.6) *in Lemma 2.1.3. But here we have a sum of multiple normal distributions, which we can not combine to one normal distribution due to the maximum or minimum, i.e.: Let us consider the clearer case of all stimuli having the same rate $\lambda > 0$. Then Equation* (2.21) *simplifies to ($\sigma = 1$)*

$$\sum_{m=1}^{M} \left[\frac{n^{(m)}}{2} \left(\varphi_s^{(m)} - \varphi_1^{(m)}\right)^2 - \sqrt{n^{(m)}} \left(\varphi_s^{(m)} - \varphi_1^{(m)}\right) z^{(m)} \right] > 0 \quad \forall\, s \neq 1,$$

with $z^{(m)} := \frac{\bar{x}^{(m)}-\varphi_1^{(m)}}{1/\sqrt{n^{(m)}}}$, which is equivalent to

$$\min_{s\neq 1} \left(\sum_{m=1}^{M} \left[\frac{n^{(m)}}{2} \left(\varphi_s^{(m)} - \varphi_1^{(m)}\right)^2 - \sqrt{n^{(m)}} \left(\varphi_s^{(m)} - \varphi_1^{(m)}\right) z^{(m)} \right] \right) > 0.$$

In order to calculate the probability to detect stimulus 1 *correctly, i.e.,*

$$p_1 = \mathbb{P}\left(\min_{s\neq 1} \left(\sum_{m=1}^{M} \left[\frac{n^{(m)}}{2} \left(\varphi_s^{(m)} - \varphi_1^{(m)}\right)^2 - \sqrt{n^{(m)}} \left(\varphi_s^{(m)} - \varphi_1^{(m)}\right) Z^{(m)} \right] \right) > 0 \right),$$

with $Z^{(1)},\dots,Z^{(M)} \sim \mathcal{N}(0,1)$ and independent, we can not combine the normal distributions, i.e.,

$$p_1 \neq \mathbb{P}\left(\min_{s\neq 1} \left(\sqrt{\sum_{m=1}^{M} n^{(m)} \left(\varphi_s^{(m)} - \varphi_1^{(m)}\right)^2} Z + \sum_{m=1}^{M} \frac{n^{(m)}}{2} \left(\varphi_s^{(m)} - \varphi_1^{(m)}\right)^2 \right) > 0 \right),$$

with $Z \sim \mathcal{N}(0,1)$, as in general for $Z_1, Z_2, Z \sim \mathcal{N}(0,1)$ and independent it holds

$$\min_{m\in\{1,2\}} \left(a_1^{(m)} Z_1 + a_2^{(m)} Z_2\right) \overset{d}{\neq} \min_{m\in\{1,2\}} \left(b^{(m)} Z\right).$$

In example we consider the case $a_1^{(1)} = a_2^{(1)} = a_1^{(2)} = 1/\sqrt{2}$ and $a_2^{(2)} = -1/\sqrt{2}$. Since both $1/\sqrt{2}(Z_1+Z_2)$ and $1/\sqrt{2}(Z_1-Z_2)$ are standard normal distributed and independent, we obtain

$$\mathbb{P}(Z_1 + Z_2 > 0 \wedge Z_1 - Z_2 > 0) = \frac{1}{4}, \quad \textit{but} \quad \mathbb{P}(Z > 0) = \frac{1}{2}.$$

In case of $S = 2$ stimuli, the detection probability can be calculated numerically efficient, for further information see the following paragraph. Otherwise we use simulations and Lemma 2.2.2.

Special case of $M = 2$ neurons and $S = 2$ stimuli

In the following we detail how to derive the detection probability numerically for two neurons and two stimuli. Similar to the case of one neuron the detection probability is maximized, if there exists one stimulus no neuron emits any spike, cf. Section 2.2.2.1. This stimulus is called nullstimulus and we do not count the nullstimulus as real stimulus. Thus we have as input a nullstimulus with rates $\lambda_0^{(1)}, \lambda_0^{(2)} = 0$ and two 'real' stimuli with rate parameters $\lambda_1^{(1)}, \lambda_1^{(2)}$ and $\lambda_2^{(1)}, \lambda_2^{(2)}$ and phase parameters $\varphi_1^{(1)}, \varphi_1^{(2)}$ and $\varphi_2^{(1)}, \varphi_2^{(2)}$.

At first, we assume that $\lambda_s^{(m)} > 0$ for all $m, s \in \{1,2\}$, and $\varphi_1^{(m)} \neq \varphi_2^{(m)}$ for all $m \in \{1,2\}$. If one neuron has equal phases for both stimuli, i.e., $\exists\, m \in \{1,2\}$ with $\varphi_1^{(m)} = \varphi_2^{(m)}$, Equation (2.21) can be easily calculated, as $\bar{X}^{(m)}$ does not matter and the calculation is mostly identical to the one neuron case. However this case will not occur in an optimal parameter set, since there will be chosen the maximal and minimal phase in every neuron.

Rearranging Equation (2.21) for $S = 2$ stimuli, the probability to detect stimulus 1 when it is present can be calculated by

$$p_1 = \mathbb{P}_1\left(\sum_{m=1}^{2}\left(g\left(N^{(m)}, \lambda_1^{(m)}, \lambda_2^{(m)}\right) - \sqrt{N^{(m)}}\left(\varphi_2^{(m)} - \varphi_1^{(m)}\right) Z^{(m)} + N^{(m)} \frac{\left(\varphi_1^{(m)} + \varphi_2^{(m)}\right)^2}{2}\right) > 0\right), \tag{2.22}$$

where $Z^{(m)} \sim \mathcal{N}(0,1)$ and $Z^{(1)}, Z^{(2)}$ are independent and

$$g\left(N^{(m)}, \lambda_1^{(m)}, \lambda_2^{(m)}\right) := N^{(m)} \log\left(\frac{\lambda_1^{(m)}}{\lambda_2^{(m)}}\right) - \left(\lambda_1^{(m)} - \lambda_2^{(m)}\right).$$

Combining both variables leads to

$$p_1 = \mathbb{P}_1\left(Z < f\left(\boldsymbol{\lambda}, \boldsymbol{\varphi}, N^{(1)}, N^{(2)}\right)\right), \tag{2.23}$$

where $Z \sim \mathcal{N}(0,1)$ and

$$f\left(\boldsymbol{\lambda}, \boldsymbol{\varphi}, N^{(1)}, N^{(2)}\right) := \frac{\sum_{m=1}^{2}\left(\lambda_2^{(m)} - \lambda_1^{(m)} + N^{(m)} \log\left(\frac{\lambda_1^{(m)}}{\lambda_2^{(m)}}\right) + \frac{1}{2} N^{(m)}\left(\varphi_1^{(m)} + \varphi_2^{(m)}\right)^2\right)}{\sqrt{N^{(1)}\left(\varphi_1^{(1)} + \varphi_2^{(1)}\right)^2 + N^{(2)}\left(\varphi_1^{(2)} + \varphi_2^{(2)}\right)^2}}.$$

Case 1: $\lambda_s^{(m)} > 0$ for all $s, m \in \{1,2\}$. Here we can directly apply Equation (2.23) and sum up the possible values of the Poisson distribution

$$p_1 = \sum_{\substack{n^{(1)},n^{(2)}=0 \\ n^{(1)}+n^{(2)}>0}}^{\infty} \frac{\left(\lambda_1^{(1)}\right)^{n^{(1)}}}{n^{(1)}!} e^{-\lambda_1^{(1)}} \frac{\left(\lambda_1^{(2)}\right)^{n^{(2)}}}{n^{(2)}!} e^{-\lambda_1^{(2)}} \mathbb{P}\left(Z < f\left(\boldsymbol{\lambda}, \boldsymbol{\varphi}, n^{(1)}, n^{(2)}\right)\right)$$

and analog for $p_2 = \mathbb{P}_2(\mathbf{B} \in A_2)$. The lower summation bound regards that in case of $n^{(1)} = n^{(2)} = 0$ we choose the nullstimulus, which yields $p_0 = 1$.

Case 2: Exactly one $\lambda_s^{(m)} = 0$ for $s, m \in \{1,2\}$, w.l.o.g. $\lambda_1^{(1)} = 0$. From Equation (2.22) and the consideration that decision for stimulus 1 is only possible, if neuron 1 produced no spikes, i.e., $n^{(1)} = 0$, we get

$$p_1 = \sum_{n^{(2)}>0}^{\infty} \frac{\left(\lambda_1^{(2)}\right)^{n^{(2)}}}{n^{(2)}!} e^{-\lambda_1^{(2)}} \mathbb{P}\left(Z > \frac{\lambda_2^{(1)} + n^{(2)} \log\left(\frac{\lambda_1^{(2)}}{\lambda_2^{(2)}}\right) - \left(\lambda_1^{(2)} - \lambda_2^{(2)}\right) + i_2 \frac{\left(\varphi_1^{(2)} - \varphi_2^{(2)}\right)^2}{2}}{\sqrt{n^{(2)}}\left(\varphi_2^{(2)} - \varphi_1^{(2)}\right)}\right)$$

for $\varphi_1^{(2)} > \varphi_2^{(2)}$, otherwise replace "$Z <$". As we always decide for stimulus 2, if neuron 1 emits a spike, Equation (2.22) yields

$$p_2 = 1 - e^{-\lambda_2^{(1)}} + \sum_{n^{(2)}>0}^{\infty}$$
$$e^{-\lambda_2^{(1)}} \frac{\left(\lambda_1^{(2)}\right)^{n^{(2)}}}{i_2!} e^{-\lambda_2^{(2)}} \mathbb{P}\left(Z < \frac{\lambda_2^{(1)} + n^{(2)} \log\left(\frac{\lambda_1^{(2)}}{\lambda_2^{(2)}}\right) - \left(\lambda_1^{(2)} - \lambda_2^{(2)}\right) - n^{(2)} \frac{\left(\varphi_1^{(2)} - \varphi_2^{(2)}\right)^2}{2}}{\sqrt{n^{(2)}}\left(\varphi_2^{(2)} - \varphi_1^{(2)}\right)}\right).$$

Case 3: Phase worthless, i.e., w.l.o.g. $\lambda_1^{(1)} = 0$ and $\lambda_2^{(2)} = 0$. There are only 3 possibilities, as both neurons do not spike for the same stimulus:

- $N^{(1)} = 0 \ \wedge \ N^{(2)} = 0 \implies$ decision for nullstimulus
- $N^{(1)} > 0 \ \wedge \ N^{(2)} = 0 \implies$ decision for stimulus 2
- $N^{(1)} = 0 \ \wedge \ N^{(2)} > 0 \implies$ decision for stimulus 1.

So we can calculate

$$p_1 = \mathbb{P}\left(N_1^{(2)} > 0\right) = 1 - e^{-\lambda_1^{(2)}} \quad \text{and} \quad p_2 = \mathbb{P}\left(N_2^{(1)} > 0\right) = 1 - e^{-\lambda_2^{(1)}}.$$

Case 4: One neuron case:

- i.e., w.l.o.g. $\lambda_1^{(1)} = 0$ and $\lambda_2^{(1)} = 0$. The detection probability can be calculated by Lemma 2.1.3, using only $\boldsymbol{\lambda}^{(2)}$ and $\boldsymbol{\varphi}^{(2)}$.

- i.e., w.l.o.g. $\lambda_1^{(1)} = 0$ and $\lambda_1^{(2)} = 0$. Here it is not possible to distinguish if stimulus 1 or the nullstimulus is present. The detection probability can be calculated by Lemma 2.1.3 using $\boldsymbol{\lambda} = \left\{0, 0, \lambda_2^{(1)} + \lambda_2^{(2)}\right\}$, since the phase is of no use, as we only distinguish one stimulus from two nullstimuli. Thus to detect stimulus 2 requires $N_2^{(1)} > 0$ or $N_2^{(2)} > 0$. Since $N_2^{(1)}$ and $N_2^{(2)}$ are independent, we know that $N_2^{(1)} + N_2^{(2)} \sim Pois\left(\lambda_2^{(1)} + \lambda_2^{(2)}\right)$.

2.2.2 Optimal parameter choices

In this section we investigate how rate and phase parameters should be chosen in order to maximize the detection probability in case of 2 neurons and various numbers of stimuli. Thereby our main goal is to quantify the increase in the detection probability of a rate and phase code compared to a pure rate code. Equivalent to Section 2.1.2 we note that p_D is not affected by a shift of the phase parameter and we can therefore assume a minimal phase parameter of zero. We also notice that p_D can always be increased by increasing the maximal rate $\lambda_M := \max_{s,m} \lambda_s^{(m)}$ and the maximal phase $\varphi_M := \max_{s,m} \varphi_s^{(m)}$. Therefore, we keep λ_M and φ_M fixed and assume each neuron can choose the same maximal rate and phase parameter, i.e., $0 \leq \lambda_s^{(m)} \leq \lambda_M$ and $0 \leq \varphi_s^{(m)} \leq \varphi_M$ for all $s = 1, \ldots, S$ and $m \in \{1, 2\}$. The results are derived as a function of these restrictions.

Similar to Section 2.1.2 we observe that p_D is maximized if both neurons have a minimal rate parameter of zero for the same stimulus, called nullstimulus, cf. Lemma 2.2.5. However the statement slightly differs from Lemma 2.1.4, as for given arbitrary parameters, it is not always optimal to decrease the rates of one stimulus to zero, see Section 2.2.2.1.

In Section 2.2.2.2 we consider $S = 2$ stimuli (plus nullstimulus) and discuss optimal parameters. Using two neurons one might ask, if more information per spike can be transmitted as in the one neuron case. Therefore, we compare a single neuron with maximal rate $2\lambda_M$ and two neurons with maximal rate λ_M. The outcome is that two neurons have a higher detection probability even for a less expected number of spikes in comparison to the one neuron case.

Another question is, if it is possible to use the optimal one dimensional solution to form the optimal or almost optimal coding in case of two neurons. Therefore, we need to calculate the optimal combination of two one dimensional optimal neurons, i.e., which parameter of the first and second neuron should be assigned to the same stimulus. However, this optimal combination has even lowered detection probability than using a single neuron with double maximal rate, see Section 2.2.2.2.

In Section 2.1.2 we found that the phase considerably increases the detection probability compared to a pure rate code, already for $S = 2$ stimuli. In case of 2 neurons and less than 4 stimuli quite large phases are needed to increase the detection probability clearly, see Section 2.2.2.2 and Section 2.2.2.3. Rate coding dominates for less than 2^M stimuli, M number of neurons, due to the stability of a binary code, cf. Figure 2.12 B. Only for a high number of stimuli already small phases increase the detection probability explicitly and thus provides additional information in a realistic parameter range (Section 2.2.2.3).
Fitting the GLO to real data out of the visual cortex of an anesthetized cat under visual stimulation, we observe phase parameters up to $\varphi_M = 0.75$ (Schneider and Nikolić, 2006),

cf. Appendix A. Choosing small rate parameters of maximal one spike per oscillation cycle for one neuron, like most stimuli are coded in the data, emphasizes another importance of the phase: For small rates ($\lambda_M \leq 1$) only the phase is able to distinguish more than $2^M - 1$ stimuli, see Section 2.2.2.3.

2.2.2.1 Nullstimulus

In case of one neuron Lemma 2.1.4 tells us that we can increase the detection probability for arbitrary parameters, if we set the minimum rate to zero. In case of two neurons we have two rate parameter vectors and in general not both neurons have their minimal rate parameter in the same stimulus. So which stimulus should be transformed in a nullstimulus and how does this choice affects the detection probability? Let us consider $S = 3$ stimuli with

$$\begin{aligned} \lambda^{(1)} &= (0, 2, 1) \quad \text{and} \quad \varphi^{(1)} = (0, 0, 10) \\ \lambda^{(2)} &= (2, 0, 1) \quad \text{and} \quad \varphi^{(2)} = (0, 0, 10), \end{aligned}$$

which yields a detection probability of $p_D \approx 0.932$. One might set $\lambda_1^{(2)} = 0$ (which is equivalent to $\lambda_2^{(1)} = 0$) to create a nullstimulus, which yields $p_D \approx 0.911$. Similar, setting $\lambda_3^{(1)} = 0$ and $\lambda_3^{(2)} = 0$ yields $p_D \approx 0.91$. Thus we can not optimize these parameter vector by setting two rate parameters to zero. However, if we create the nullstimulus by rotating the rate parameters of neuron 2, i.e.,

$$\begin{aligned} \lambda^{(1)} &= (0, 2, 1) \quad \text{and} \quad \varphi^{(1)} = (0, 0, 10) \\ \lambda^{(2)} &= (0, 2, 1) \quad \text{and} \quad \varphi^{(2)} = (0, 0, 10), \end{aligned}$$

we can increase the detection probability and obtain $p_D \approx 0.948$. In general, we will see in Lemma 2.2.5, that parameter vectors, which maximize the detection probability, always have a nullstimulus. The basic idea is shown in Figure 2.12 A.

Lemma 2.2.5. *Consider S stimuli and M neurons and given maximal rate $\lambda_M \geq 0$ and maximal phase $\varphi_M \geq 0$. We can find parameters $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^{(1)}, \dots, \boldsymbol{\lambda}^{(M)})$ and $\boldsymbol{\varphi} = (\boldsymbol{\varphi}^{(1)}, \dots, \boldsymbol{\varphi}^{(M)})$ with $\exists\, s \in \{1, \dots, S\} : \lambda_s^{(m)} = 0 \,\forall m \in \{1, \dots, M\}$ that are optimal, i.e.,*

$$p_D(\boldsymbol{\lambda}, \boldsymbol{\varphi}) = \max_{\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\varphi}}} p_D(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\varphi}}).$$

Proof. We recall that $p_D = 1/S(p_1 + \cdots + p_S)$, with p_s denoting the detection probability of stimulus s. Let w.l.o.g. stimulus 1 be the nullstimulus, i.e., $\lambda_1^{(m)} = 0 \,\forall m \in \{1, \dots, M\}$, and there exists no other nullstimulus. According to Corollary 2.2.3 we decide for stimulus 1 only if $N^{(1)} = \cdots = N^{(M)} = 0$, thus $p_1 = 1$. We use only one possible realization to decide for stimulus 1 and get the maximal possible detection probability in that realization ($\mathbb{P}_1(N^{(1)} = \cdots = N^{(M)} = 0) = 1$). All other realizations remain for the other stimuli, cf. Figure 2.12 for $S = 2$ stimuli and $M = 2$ neurons, what maximizes the detection probabilities for the other stimuli.
Since the phase is of no use if $N^{(1)} = \cdots = N^{(M)} = 0$, the possible φ_M does not affect the optimality of a nullstimulus. □

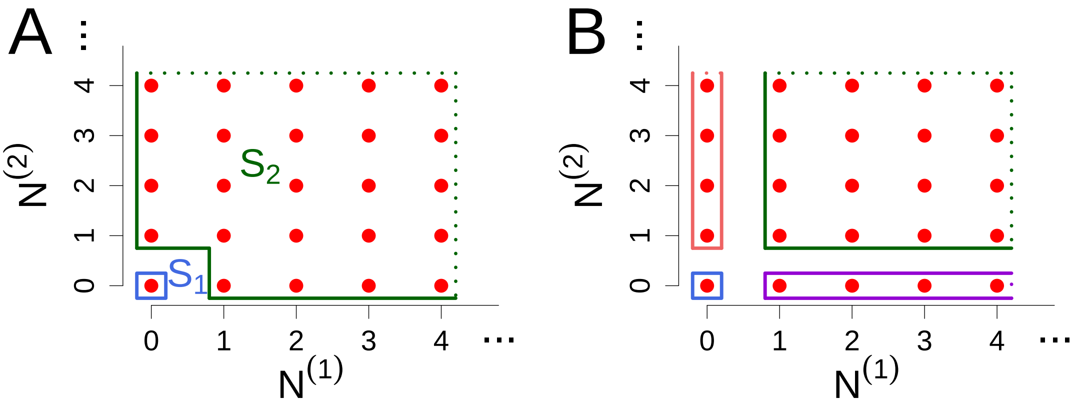


Figure 2.12: Pure rate coding for $M = 2$ neurons and $S = 2$ (A) and $S = 4$ (B) stimuli. The red dots represent all possible realizations. The dotted lines demonstrate, that the decision areas continue in this direction. A. Illustration of the optimality of a nullstimulus (S_1, blue). We only decide for S_1 if both neurons emit no spike (blue rectangle). All other possible numbers of spikes (green box) remain for stimulus 2. B. Illustration of the binary coding: We have a nullstimulus (blue), we decide in case of no spikes, a stimulus only neuron 1 emits spikes (pink), a stimulus only neuron 2 emits spikes (red) and a stimulus both neurons emit spikes (green). The decision areas are shown as rectangles.

2.2.2.2 Two stimuli

Here we consider two stimuli and discuss optimal rate and phase parameters. For $S = 2$ stimuli we need to choose 8 parameters. Using the results of Paragraph 'Special case of $M = 2$ neurons and $S = 2$ stimuli' we can numerically optimize the detection probability. The structure of the optimal coding pattern can be written as

$$\begin{aligned} &\textbf{rate: } \boldsymbol{\lambda}^{(1)} = \left(\lambda_1^{(1)}, \lambda_M\right) \text{ and } \boldsymbol{\lambda}^{(2)} = \left(\lambda_M, \lambda_2^{(2)}\right) \quad \text{for } \lambda_1^{(1)}, \lambda_2^{(2)} \in [0, \lambda_M], \\ &\textbf{phase: } \varphi_1^{(m)} - \varphi_2^{(m)} = \varphi_M \text{ for } m \in \{1, 2\}. \end{aligned}$$

where the values of $\lambda_1^{(1)}$ and $\lambda_2^{(2)}$ depend on φ_M. The precise mapping of the phase parameters does not affect the detection probability, only the maximum phase difference has to be chosen in each neuron. The minimal rate parameters are not assigned to the same stimulus, as otherwise it is more likely to falsely decide for the nullstimulus.

The exact progress of the rate parameters is shown in Figure 2.13 A for $\lambda_M = 1$ and B for $\lambda_M = 2$. In both cases it is optimal that there exists a stimulus both neurons are responding with spikes, green line. Due to the small rate parameters, the probability to emit no spike is not negligible. So this coding pattern minimizes false decisions for the nullstimulus. For $\lambda_M = 2$ and small phases it is even better to distinguish in the rate parameter and choose a medium rate of 1 for this stimulus, Figure 2.13 B green line.

For a wide parameter range a binary coding pattern is optimal, i.e., either only one neuron or both neurons react. So here the phase only supports the rate decision. For this coding pattern the phase can not increase the detection probability significantly, see Figure 2.13 C and D, blue line compared to red dotted line. In case of quite high phases, $\varphi_M \geq 1$ (A) or $\varphi_M \geq 1.4$ (B) a pure phase is optimal and the detection probability is significantly increased compared to a pure rate code (Figure 2.13 C and D).

However, in case of a pure phase code two neurons have the same detection probability as a single neuron with double maximal rate ((Figure 2.13 C and D, blue and green line): As both neurons react independent to stimulus s and the mapping of φ does not affect the detection probability (w.l.o.g. $\varphi_s^{(1)} = \varphi_s^{(2)}$), both neurons follow an independent inhomogeneous Poisson processes with intensity function $\lambda_M \phi_{\varphi_s^{(m)}, \sigma^2/n^{(m)}}(t)$, where $\phi_{\varphi_s^{(m)}, \sigma^2/n^{(m)}}$ is the density of the normal distribution with mean $\varphi_s^{(m)}$ and variance $\sigma^2/n^{(m)}$. So both processes can be combined to one inhomogeneous Poisson processes with intensity function $2\lambda_M \phi_{\varphi_s^{(m)}, \sigma^2/n^{(m)}}(t)$, the single neuron case.

In Section 2.1.2.3 the optimal rate and phase parameters are computed for two stimuli and a single neuron. Now we want to compare two neurons which are optimal for the setting of a single neuron with the optimal solution of two neurons. Therefore we need to specify how to combine both single neurons. Again the minimal rate parameters are not assigned to the same stimulus, so this case is labeled "2 cross neurons" (Figure 2.13 C and D, purple line). Two one dimensional optimal neurons code even less information as a single neuron with double maximal rate. Thus the one dimensional solution can not be used to create an acceptable two-dimensional solution.

In the previous analyses the two neuron case is compared to one neuron with double maximal rate. To assess whether the two-dimensional solution codes information more efficiently, the mean firing rate is compared in Figure 2.13 E and F. Except the case of $\lambda_M = 2$ and $\varphi_M \geq 0.5$ (Figure 2.13 F) two optimal neurons need significantly less spikes to code even more information.

2.2.2.3 Three and four stimuli

In this section we want to evaluate if the phase can provide additional information in case of $M = 2$ neurons and $S = 3$ or $S = 4$ stimuli. Therefore we first determine numerically the optimal rate code, using Corollary 2.2.3. As in case of a rate and phase code the detection probability can be calculated only by simulation for $S \geq 3$ stimuli, see Lemma 2.2.2, and the optimization is quite expensive due to the high number of parameters ($2S$ rate and $2S$ phase parameters), we investigate parameter combinations, which are plausible according to our results of a pure rate code and the one neuron case.

So the following analyzed coding patterns should be rather considered as option to quantify the impact of the phase on correct detection. However, the presented coding patterns are confirmed by simulations and coincide with our findings so far.

Simulations have shown that a binary coding dominates especially for small phases. Why we use the term binary code is illustrated in Figure 2.12 B. Three stimuli (plus nullstimulus) can be coded quite efficiently with a pure rate code. If we simultaneously consider different phases in a binary code, the phase can only support a more precise stimulus decision within the colored decision areas. For example let us consider the rate parameters $\boldsymbol{\lambda}^{(1)} = (\lambda_M, 0, \lambda_M)$ and $\boldsymbol{\lambda}^{(2)} = (0, \lambda_M, \lambda_M)$, where stimulus 3 corresponds to the green area and stimulus 1 to

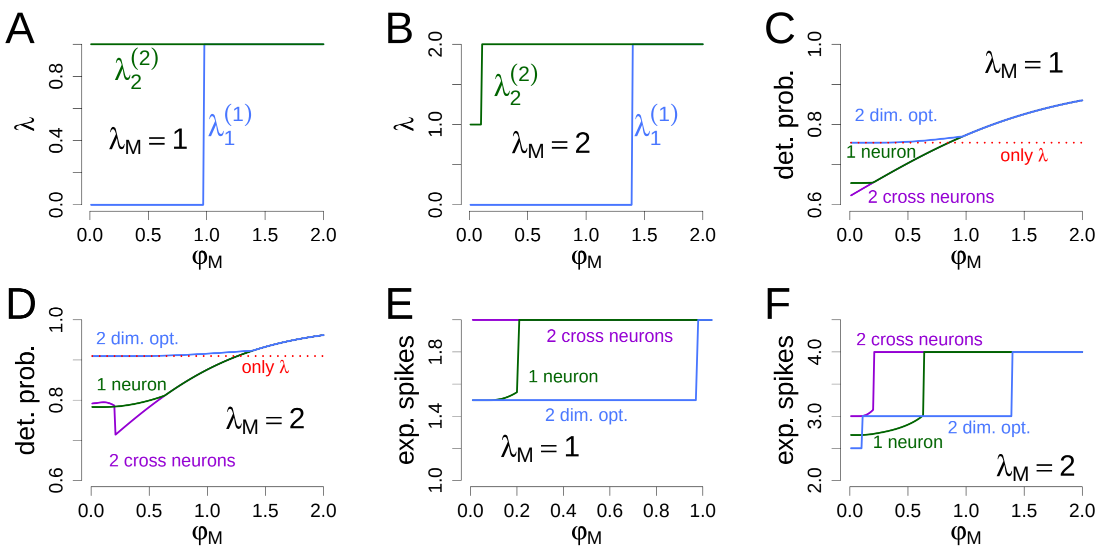


Figure 2.13: A and B. Optimal rate parameters for two stimuli and 2 neurons. The optimal coding pattern is $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\varphi}^{(1)}) = (\lambda_1^{(1)}, \lambda_M, 0, \varphi_M)$ and $(\boldsymbol{\lambda}^{(2)}, \boldsymbol{\varphi}^{(2)}) = (\lambda_M, \lambda_2^{(2)}, 0, \varphi_M)$ (the phase mapping does not matter). The optimal $\lambda_1^{(1)}$ and $\lambda_2^{(2)}$ is shown as a function of φ_M for $\lambda_M = 1$ (A) and $\lambda_M = 2$ (B). C ($\lambda_M = 1$) and D ($\lambda_M = 2$). Detection probability as a function of φ_M for three cases: 1) Optimal solution for two neuron ("2 dim. opt."), 2) a single neuron with double maximal rate ("1 neuron"), 3) optimal combination of one dimensional optimal neurons ("2 cross neurons"). The decrease of the detection probability in the case of "2 cross neurons" for $\lambda_M = 2$ is due to the jump in a pure phase code in the one dimensional case. The optimal rate code in case of two neurons is shown as red dotted line. E and F. Mean firing rate for the three cases. For almost all phases two neurons code information verifiably more efficient.

the purple area, cf. Figure 2.12 B. Assume stimulus 3 is present, but we observe the unlikely event $N^{(2)} = 0$. If $N^{(1)} > 0$ we would decide falsely for stimulus 1 (purple), if $N^{(1)} = 0$ we would decide falsely for the nullstimulus (blue). Let us assume $N^{(1)} > 0$. If we can additional use the phase to distinguish stimulus 1 and 3 in neuron 1, i.e., $\varphi_1^{(1)} = 0$ and $\varphi_1^{(3)} = \varphi_M$, it is possible that we decide correctly for stimulus 3, for example if $\bar{x}^{(1)}$ is quite large. However, our phases are to small to correct a false decision due to the rate in many simulations.
For $S = 3$ stimuli we compare the following three coding patterns:

$$\begin{array}{rll}
\textbf{rate based:} & \boldsymbol{\lambda}^{(1)} = (\lambda_M, 0, \lambda_M) & \boldsymbol{\varphi}^{(1)} = (0, 0, \varphi_M) \\
& \boldsymbol{\lambda}^{(2)} = (0, \lambda_M, \lambda_M) & \boldsymbol{\varphi}^{(2)} = (0, 0, \varphi_M) \\
\textbf{rate and phase based:} & \boldsymbol{\lambda}^{(1)} = (\lambda_M, \lambda_M, \lambda_M) & \boldsymbol{\varphi}^{(1)} = (0, \varphi_M/2, \varphi_M) \\
& \boldsymbol{\lambda}^{(2)} = (\lambda_M, 0, \lambda_M) & \boldsymbol{\varphi}^{(2)} = (0, 0, \varphi_M) \\
\textbf{only phase:} & \boldsymbol{\lambda}^{(1)} = (\lambda_M, \lambda_M, \lambda_M) & \boldsymbol{\varphi}^{(1)} = (0, \varphi_M/2, \varphi_M) \\
& \boldsymbol{\lambda}^{(2)} = (\lambda_M, \lambda_M, \lambda_M) & \boldsymbol{\varphi}^{(2)} = (\varphi_M/2, 0, \varphi_M).
\end{array}$$

The 'rate based' code corresponds to a binary code, but is supported by the phase. Stimuli with

the same rate parameters (within one neuron) are distinguished by maximal phase difference. Only rate (the optimal rate code) is the 'rate based' code, but $\varphi_M = 0$.
In case of a 'rate and phase based' code two stimuli exist which both neurons have maximal firing rate and the phase is needed to distinguish these two stimuli. In case of a 'only phase' code, both neurons code with a pure phase code. The parameters are connected such that stimuli, which are difficult to distinguish in one neuron, are separated more strictly in the second neuron.
In Figure 2.14 A (C) the detection probability of the three cases is compared for $\lambda_M = 1$ ($\lambda_M = 2$). We need quite large phases to increase the detection probability compared to a pure rate code (red dotted line). Even for relative high phases the rate based coding pattern (red line) maximizes the detection probability among the other patterns (blue and green). However, in most part of the considered parameter range the rate code dominates obviously and the phase can increase the detection probability only a little.

Different for $S = 4$ stimuli: Here the rate is not able to distinguish all stimuli for maximal rate $\lambda_M \leq 1$, as with $\lambda_M \leq 1$ we can only construct the 3 decision areas (plus nullstimulus), which are shown in Figure 2.13 B. An additional stimulus with rate $0 < \lambda < \lambda_M$ will never be detected as it has never maximal weight, cf. Corollary 2.2.3. In Figure 2.14 B (green line) we compare the following coding pattern

$$\begin{aligned} \textbf{rate-3-phase-2:}\quad & \boldsymbol{\lambda}^{(1)} = (0, \lambda_M, \lambda_M, \lambda_M) \quad \boldsymbol{\varphi}^{(1)} = (0, 0, 0, \varphi_M) \\ & \boldsymbol{\lambda}^{(2)} = (\lambda_M, 0, \lambda_M, \lambda_M) \quad \boldsymbol{\varphi}^{(2)} = (0, 0, 0, \varphi_M) \end{aligned}$$

to a pure rate code (optimal rate code), i.e., the same rate parameters, but $\varphi_M = 0$. Here a phase of $\varphi_M = 0.75$ increases the detection probability evidently (green line vs red dotted line) and provides additional information, as the rate can not distinguish stimulus 3 and 4. The purple line indicates again that one neuron with a maximal rate of $\lambda_M = 2$ is no alternative to two neurons with $\lambda_M = 1$.

But also in case of a higher rate $\lambda_M = 2$, where a pure rate code can distinguish all stimuli, only small phase are needed (green or blue line) to significantly increase the detection probability in comparison to an optimal rate code (red dotted line), see Figure 2.14 D. For $S = 4$ stimuli the detection probability is further increased by the phase for an increasing maximal rate λ_M due to the higher accuracy of the mean spike time, cf. Figure 2.14 B and D.
For $S = 4$ stimuli and maximal rate $\lambda_M = 2$ we compare the following four coding patterns

$$\begin{aligned} \textbf{rate-4-phase-2:}\quad & \boldsymbol{\lambda}^{(1)} = (\lambda_M, 0, \lambda, \lambda_M) && \boldsymbol{\varphi}^{(1)} = (0, 0, 0, \varphi_M) \\ & \boldsymbol{\lambda}^{(2)} = (0, \lambda_M, \lambda, \lambda_M) && \boldsymbol{\varphi}^{(2)} = (0, 0, 0, \varphi_M) \\ \textbf{rate-4-phase-3:}\quad & \boldsymbol{\lambda}^{(1)} = (\lambda_M, 0, \lambda, \lambda_M) && \boldsymbol{\varphi}^{(1)} = (0, 0, \varphi_M/2, \varphi_M) \\ & \boldsymbol{\lambda}^{(2)} = (0, \lambda_M, \lambda, \lambda_M) && \boldsymbol{\varphi}^{(2)} = (0, 0, \varphi_M/2, \varphi_M) \\ \textbf{rate-3-phase-2:}\quad & \boldsymbol{\lambda}^{(1)} = (0, \lambda_M, \lambda_M, \lambda_M) && \boldsymbol{\varphi}^{(1)} = (0, 0, 0, \varphi_M) \\ & \boldsymbol{\lambda}^{(2)} = (\lambda_M, 0, \lambda_M, \lambda_M) && \boldsymbol{\varphi}^{(2)} = (0, 0, 0, \varphi_M) \\ \textbf{rate-3-phase-3:}\quad & \boldsymbol{\lambda}^{(1)} = (0, \lambda_M, \lambda_M, \lambda_M) && \boldsymbol{\varphi}^{(1)} = (0, \varphi_M/2, 0, \varphi_M) \\ & \boldsymbol{\lambda}^{(2)} = (\lambda_M, 0, \lambda_M, \lambda_M) && \boldsymbol{\varphi}^{(2)} = (\varphi_M/2, 0, 0, \varphi_M). \end{aligned}$$

For $\lambda_M = 2$ the optimal rate code in 'rate-4-phase-2' or 'rate-4-phase-3' is given by $\lambda = 1$. It should be noted that for large phases ($\varphi_M \approx 2$), the detection probability is increased (a little) by switching the rate parameter λ and λ_M in one neuron (in 'rate-4-phase-2' and 'rate-4-phase-3').

Naturally we would expect that the parameter set 'rate-4-phase-2' dominates for small phases, but already for a maximal phase $\varphi \approx 0.2$ parameter set 'rate-3-phase-2' yields a higher detection probability, cf. Figure 2.14 D, red line vs green line. Larger phases result in a more detailed phase code, i.e., it is advantageous to implement a middle phase value ('rate-3-phase-3', blue line).

To summarize, due to the stability of a binary rate code, small phases can increase the detection probability compared to a pure rate code only for $S \geq 2^M$ Stimuli (M number of neurons) or quite large phases are needed.

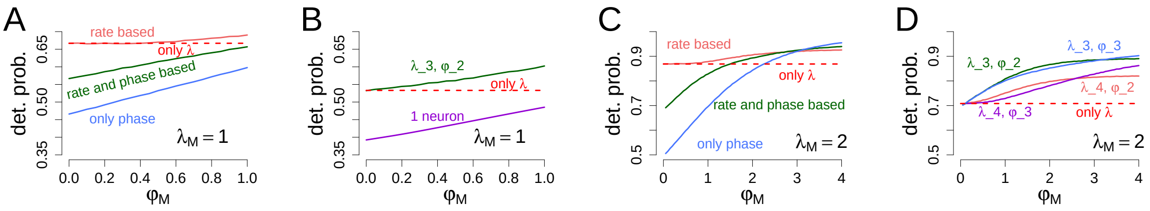


Figure 2.14: A and C. Detection probability of different coding patterns as a function of φ_M for three stimuli and $\lambda_M = 1$ (A) and $\lambda_M = 2$ (C): 'rate based' $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\varphi}^{(1)}) = ((\lambda_M, 0, \lambda_M), (0, 0, \varphi_M))$ and $(\boldsymbol{\lambda}^{(2)}, \boldsymbol{\varphi}^{(2)}) = ((0, \lambda_M, \lambda_M)), (0, 0, \varphi_M))$, 'rate and phase based' $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\varphi}^{(1)}) = ((\lambda_M, \lambda_M, \lambda_M), (0, \varphi_M/2, \varphi_M))$ and $(\boldsymbol{\lambda}^{(2)}, \boldsymbol{\varphi}^{(2)}) = ((\lambda_M, 0, \lambda_M)), (0, 0, \varphi_M))$, 'only phase' $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\varphi}^{(1)}) = ((\lambda_M, \lambda_M, \lambda_M), (0, \varphi_M/2, \varphi_M))$ and $(\boldsymbol{\lambda}^{(2)}, \boldsymbol{\varphi}^{(2)}) = ((\lambda_M, \lambda_M, \lambda_M)), (\varphi_M/2, 0, \varphi_M))$. B and D. Detection probability of different coding patterns as a function of φ_M for four stimuli and $\lambda_M = 1$ (B) and $\lambda_M = 2$ (D). B. We consider 'rate-3-phase-2' $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\varphi}^{(1)}) = ((0, \lambda_M, \lambda_M, \lambda_M), (0, 0, 0, \varphi_M))$ and $(\boldsymbol{\lambda}^{(2)}, \Phi^2) = ((\lambda_M, 0, \lambda_M, \lambda_M), (0, 0, 0, \varphi_M))$. '1 neuron' illustrates the maximal detection probability using only one neuron but double maximal rate. D. We consider 'rate-4-phase-2' $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\varphi}^{(1)}) = ((\lambda_M, 0, \lambda, \lambda_M), (0, 0, 0, \varphi_M))$ and $(\boldsymbol{\lambda}^{(2)}, \boldsymbol{\varphi}^{(2)}) = ((0, \lambda_M, \lambda, \lambda_M), (0, 0, 0, \varphi_M))$, 'rate-4-phase-3' $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\varphi}^{(1)}) = ((\lambda_M, 0, \lambda, \lambda_M), (0, 0, \varphi_M/2, \varphi_M))$ and $(\boldsymbol{\lambda}^{(2)}, \boldsymbol{\varphi}^{(2)}) = ((0, \lambda_M, \lambda, \lambda_M), (0, 0, \varphi_M/2, \varphi_M))$, 'rate-3-phase-2' $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\varphi}^{(1)}) = ((0, \lambda_M, \lambda_M, \lambda_M), (0, 0, 0, \varphi_M))$ and $(\boldsymbol{\lambda}^{(2)}, \boldsymbol{\varphi}^{(2)}) = ((\lambda_M, 0, \lambda_M, \lambda_M), (0, 0, 0, \varphi_M))$, 'rate-3-phase-3' $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\varphi}^{(1)}) = ((0, \lambda_M, \lambda_M, \lambda_M), (0, \varphi_M/2, 0, \varphi_M))$ and $(\boldsymbol{\lambda}^{(2)}, \boldsymbol{\varphi}^{(2)}) = ((\lambda_M, 0, \lambda_M, \lambda_M), (\varphi_M/2, 0, 0, \varphi_M))$. A, B, C and D. The optimal rate code is shown as red dashed line.

Chapter 3

Overview Bayesian Inference

In the following sections we give a short introduction to Bayesian inference and an overview of general known results, which we later make use of in application of the Bayesian Online Change Point Detection Algorithm (BOCD). Therefore, we mainly follow the book of Gelman et al. (2013), but give a more in-depth insight, especially in exponential family distributions (Brown, 1986; Consonni and Veronese, 1992; Gutiérrez-Pena and Smith, 1995, 2003) and conjugate prior distributions (Diaconis and Ylvisaker, 1979, 1985). Further references for the theory of Bayesian inference are Ghosh and Ramamoorthi (2003); Schervish (1995).
We start with a short formalization of Bayesian Inference in Section 3.1 and an interpretation of Bayes formula in a discrete or continuous or mixed distribution setting, see Section 3.1.1. Afterwards in Section 3.2 we practical introduce to Bayesian inference by the example of a coin toss with random success probability and motivate the main theoretic results about posterior prediction.
One important theoretic result is that as long as the prior distribution assigns some weight to the true parameter value, we have posterior consistency, see Section 3.3. Under some regulatory conditions the posterior distribution even satisfies asymptotic normality with the maximum likelihood estimator as expectation.
Later we are especially interested in the application of a Bayesian Online Change point algorithm, where a computationally efficient computation of the posterior and predictive distribution is necessary. In Section 3.4 we formalize the concept of conjugacy and see that in case of an exponential family distribution a standard conjugate prior distribution exists and the predictive distribution can be determined analytically. Also, the useful property of posterior linearity in the expectation of the sufficient statistic holds in general for exponential family distribution and its standard conjugate prior.

3.1 Bayesian model

Contrary to the classical statistic in the Bayesian statistic the unknown parameter θ of a distribution P_θ of a data set is understood itself as a realization of a random variable Θ. Thus it is reasonable to express our belief about Θ with probabilities. Before observing the data, we start with a prior distribution of θ, which represents our start knowledge. After observing the data, we use Bayes rule to update our belief about Θ in light of new information and call the outcome posterior distribution. This process is basically referred to as Bayesian inference. In the following we formalize the Bayesian approach and introduce the used notations.

Definition 3.1.1. *Let $\mathcal{P} = \{P_\theta : \theta \in \Omega_\theta\}$ be a family of probability distributions, where Ω_θ denotes the parameter space. In a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ the parameter Θ is drawn randomly, i.e.,*

$$\Theta \sim \Pi,$$

where Π denotes the distribution of Θ, and

$$X_1, \ldots, X_k \,|\, \{\Theta = \theta\} \sim P_\theta, \ k \geq 1,$$

where given $\{\Theta = \theta\}$ $X_1, \ldots, X_k$ are independent and have conditional distribution P_θ. In short we notate $X_1 \sim P_\Theta$.

Remark 3.1.2. *P_θ is called sample distribution and the conditional weights or the conditional density of X are denoted by $p_\theta(\cdot)$. We consider only the case $\Omega_\theta \subseteq \mathbb{R}^d$.*

Definition 3.1.3. *In a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ the distribution Π of Θ is called prior distribution.*

Remark 3.1.4. *We always assume that the prior distribution Π has a continuous density function $\pi(\cdot)$, which should be called prior density function. However, in Bayesian inference the terms 'distribution' and ' density' are used interchangeably. To keep notation compact we partly follow the Bayesian language.*

In Bayesian inference we want to use the information of realization x of X to create a more precise knowledge about the realization of the unknown parameter θ. Thus we are interested in the conditional distribution of Θ given $\{X = x\}$.

Definition 3.1.5. *In a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ the conditional distribution of Θ given $\{X = x\}$, denoted as $\Pi \,|\, \{X = x\}$, is called posterior distribution. Analogously the posterior density is denoted by $\pi(\cdot \,|\, X = x)$.*

Definition 3.1.6. *In a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ the unconditional distribution of X, denoted by $p(\cdot)$, is called prior predictive distribution.*

Definition 3.1.7. *In a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ the conditional distribution of X_{k+1} given $\{X_1 = x_1, \ldots, X_k = x_k\}$, denoted by $p(\cdot \,|\, X_{1:k} = x_{1:k})$, is called predictive distribution.*

3.1.1 Bayes formula

To calculate the posterior distribution, we use Bayes formula. As we consider both cases: X is a discrete random variable with conditional weights $p_\theta(\cdot)$ and X is a continuous random variable with conditional density $p_\theta(\cdot)$, we review Bayes formula for the possible combinations of Θ and X in Sections 3.1.1.1 to 3.1.1.3.

3.1.1.1 The discrete case

Here we consider the case that both X and Θ are discrete random variables, to motivate Bayes formula in the originally setting.

Definition 3.1.8 (Conditional probability)**.** *Let X and Θ be discrete random variables. Then the conditional probability of $\{\Theta = \theta\}$ given $\{X = x\}$ is defined as*

$$\mathbb{P}(\Theta = \theta \,|\, X = x) := \frac{\mathbb{P}(\Theta = \theta, X = x)}{\mathbb{P}(X = x)}.$$

If $\mathbb{P}(X = x) = 0$, then $\mathbb{P}(\Theta = \theta \,|\, X = x) := 0$.

Lemma 3.1.9 (Law of total probability)**.** *Let X and Θ be discrete random variables with $\mathbb{P}(\Theta \in \Omega_\theta) = 1$. Then*

$$\mathbb{P}(X = x) = \sum_{\theta \in \Omega_\theta} \mathbb{P}(X = x \,|\, \Theta = \theta) \cdot \mathbb{P}(\Theta = \theta).$$

Proof. Reverting the definition of conditional probability yields

$$\begin{aligned}\mathbb{P}(X = x) &= \mathbb{P}(X = x, \Theta \in \Omega_\theta)\\ &= \sum_{\theta \in \Omega_\theta} \mathbb{P}(X = x, \Theta = \theta) = \sum_{\theta \in \Omega_\theta} \mathbb{P}(X = x \,|\, \Theta = \theta)\mathbb{P}(\Theta = \theta).\end{aligned}$$

□

Lemma 3.1.10 (Bayes formula - discrete)**.** *Let X and Θ be discrete random variables with $\mathbb{P}(\Theta \in \Omega_\theta) = 1$. Then the conditional probability of $\{\Theta = \theta\}$ given $\{X = x\}$ can be calculated by*

$$\mathbb{P}(\Theta = \theta \,|\, X = x) = \frac{\mathbb{P}(X = x \,|\, \Theta = \theta) \cdot \mathbb{P}(\Theta = \theta)}{\sum_{\tilde{\theta} \in \Omega_\theta} \mathbb{P}(X = x \,|\, \Theta = \tilde{\theta}) \cdot \mathbb{P}(\Theta = \tilde{\theta})}.$$

Proof. Applying the definition of conditional probability to the numerator backwards and the law of total probability to the denominator yields the statement. □

3.1.1.2 The continuous case

Here we consider the case that both X and Θ are continuous random variables with densities $p_\theta(\cdot)$ and $\pi(\cdot)$.

Definition 3.1.11 (Conditional density)**.** *Let $f(\cdot,\cdot)$ be the joint density of Θ and X and $p(\cdot)$ the density of X. Then the conditional density of Θ given $\{X = x\}$ is defined as*

$$\pi(\theta \,|\, X = x) := \frac{f(\theta, x)}{p(x)}.$$

If $p(x) = 0$, then $\pi(\theta \,|\, X = x) := 0$.

Remark 3.1.12. *In a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ with prior density $\pi(\cdot)$ and conditional density $p_\theta(\cdot)$ there exist a joint density and is given by*

$$f_{\theta,x}(x, \theta) = p_\theta(x)\pi(\theta),$$

as

$$\int_\theta \int_x p_\theta(x)\pi(\theta)\, dx\, d\theta = \int_\theta \pi(\theta) \int_x p_\theta(x)\, dx\, d\theta = \int_\theta \pi(\theta) 1\, d\theta = 1.$$

Claim 3.1.13. *Consider a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ with prior density $\pi(\cdot)$ and conditional density $p_\theta(\cdot)$. The prior predicitive distribution $p(\cdot)$ is given by*

$$p(x) = \int_\theta p_\theta(x)\pi(\theta)\, d\theta.$$

Proof. The Claim follows directly from Remark 3.1.12 and the law of total probability. □

Lemma 3.1.14 (Bayes formula - continuous)**.** *Let X be a random variable with conditional density $p_\theta(\cdot)$ and Θ a random variable with density $\pi(\cdot)$. Then the conditional density of Θ given $\{X = x\}$ can be calculated by*

$$\pi(\theta \,|\, X = x) = \frac{p_\theta(x) \cdot \pi(\theta)}{\int_{\tilde{\theta}} p_{\tilde{\theta}}(x) \cdot \pi(\tilde{\theta})}.$$

Proof. Applying the definition of conditional density to the numerator backwards and Claim 3.1.13 to the denominator yields the statement. □

Claim 3.1.15. *In a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ with prior density $\pi(\cdot)$ and conditional density $p_\theta(\cdot)$ the predictive density of X_2 given $\{X_1 = x_1\}$ can be calculated by*

$$p(x_2 \,|\, x_1) = \int_\theta p_\theta(x_2) \frac{\pi(\theta) p_\theta(x_1)}{p(x_1)}\, d\theta.$$

Proof. The Claim follows directly from the Law of total probability and Lemma 3.1.14. □

3.1.1.3 The mixed case

Here we consider the case that X is a discrete random variable with conditional weights $p_\theta(\cdot)$, but Θ is a continuous random variable with density $\pi(\cdot)$. Thereby we have the problem, that no well-defined joint density of X and Θ exists and we can not define the conditional density of Θ given X as above. Thus we consider the conditional distribution function:

Definition 3.1.16. *Let X be a discrete and Θ be a continuous random variable. Then the conditional cumulative distribution function is of the form*

$$F_{\Theta \,|\, X}(\theta \,|\, X = x) := \mathbb{P}(\Theta \leq \theta \,|\, X = x)$$

and the conditional probability density function is of the form

$$\pi(\theta \,|\, X = x) := \frac{dF_{\Theta \,|\, X}(\theta \,|\, X = x)}{d\theta}.$$

Claim 3.1.17. *In a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ with prior density $\pi(\cdot)$ and conditional weights $p_\theta(\cdot)$ the conditional cumulative distribution function can be calculated by*

$$F_{\Theta \,|\, X}(\theta \,|\, X = x) = \frac{\int_{-\infty}^{\theta} \pi(\tilde{\theta}) p_{\tilde{\theta}}(x)\, d\tilde{\theta}}{\mathbb{P}(X = x)},$$

with

$$\mathbb{P}(X = x) = \int_{\tilde{\theta}} \pi(\tilde{\theta}) p_{\tilde{\theta}}(x)\, d\tilde{\theta}.$$

Proof. The Law of total expectation yields

$$\mathbb{P}(X = x) = \mathbb{E}[\mathbb{1}_{\{X=x\}}] = \mathbb{E}\left[\mathbb{E}[\mathbb{1}_{\{X=x\}} \,|\, \Theta]\right] = \int_\theta \pi(\theta)p_\theta(x)\,d\theta$$

and analogous for $\mathbb{P}(\Theta \leq \theta, X = x)$. □

Remark 3.1.18. *As the conditional distribution function is obtained by integrating $\pi(\theta)p_\theta(x)$ (and $p(x) = \mathbb{P}(X = x)$ is a constant), the conditional density $\pi(\theta \,|\, X = x)$ exists and is proportional to $\pi(\theta)p_\theta(x)$. However, some could have the idea to call $\pi(\theta)p_\theta(x)$ the joint density of X and Θ, even it is not well-defined, but it fulfills the property*

$$\mathbb{P}(\Theta \in A, X \in B) = \int_{\theta \in A} \left(\sum_{x \in B} \pi(\theta)p_\theta(x) \right) d\theta.$$

To illustrate this approach, we give the example of a coin toss with uniform distributed success probability:

Example 3.1.19. *Consider a Bayesian model with $\Theta \sim \mathcal{U}nif(0,1)$ and $X \sim Ber(\Theta)$. The marginal distribution can be calculated by*

$$\mathbb{P}(X = 1) = \int_\theta p_\theta(1)\pi(\theta)\,d\theta = \int_0^1 \theta\,d\theta = 1/2.$$

With that we can determine the conditional distribution function

$$F_{\Theta \,|\, X}(\theta \,|\, X = x) = \begin{cases} \theta^2, & \text{for } x = 1 \\ 1 - (1-\theta)^2, & \text{for } x = 0, \end{cases}$$

as

$$\mathbb{P}(\Theta \leq \theta \,|\, X = 1) = 2\int_0^\theta \tilde{\theta}\,d\tilde{\theta} = \theta^2$$

and

$$\mathbb{P}(\Theta \leq \theta \,|\, X = 0) = 2\int_0^\theta (1 - \tilde{\theta})\,d\tilde{\theta} = 1 - (1-\theta)^2.$$

The conditional density of Θ given $\{X = x\}$ can be obtained by derivation

$$\pi(\theta \,|\, X = x) = \begin{cases} 2\theta, & \text{for } x = 1 \\ 2(1-\theta), & \text{for } x = 0, \end{cases}$$

and is as mentioned in Remark 3.1.18 proportional to $\pi(\theta)p_\theta(x) = p_\theta(x)$. Informally we can call

$$\pi(\theta)p_\theta(x) = \begin{cases} \theta, & \text{for } x = 1 \\ (1-\theta), & \text{for } x = 0 \end{cases}$$

the 'joint density' of Θ and X.

With Claim 3.1.17 and Remark 3.1.18 we can state Bayes formula in the mixed case:

Lemma 3.1.20 (Bayes formula - mixed)**.** *Let X be a discrete random variable with weights $p(\cdot)$ and Θ a random variable with density $\pi(\cdot)$. Then the conditional density of Θ given $\{X = x\}$ can be calculated by*

$$\pi(\theta \,|\, X = x) = \frac{p_\theta(x) \cdot \pi(\theta)}{\int_{\tilde{\theta}} p_{\tilde{\theta}}(x) \cdot \pi(\tilde{\theta})}.$$

Remark 3.1.21. *Bayes formula is mostly written as a statement of densities, but as original stated, it is beneficial to think about Bayes as a statement of probabilities.In case of a discrete sampling distribution and continuous prior distribution we define*

$$\begin{aligned} B :=& \{\Theta \text{ is in an interval of width } d\theta \text{ around the value } \theta\}, \\ A :=& \{X = x\}, \end{aligned}$$

which yields

$$\pi(\theta \,|\, X = x)\, d\theta = \mathbb{P}(B \,|\, A) = \frac{\mathbb{P}(A \,|\, B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{p_\theta(x) \cdot \pi(\theta)\, d\theta}{p(x)}.$$

However, as $d\theta$ is in both numerators, it is usually omitted.

Remark 3.1.22. *Bayes formula is often notated as*

$$\pi(\theta \,|\, X = x) \sim p_\theta(x)\pi(\theta),$$

which in words mean: the posterior distribution is proportional to the product of the prior distribution and the likelihood. So the above terms are equal up to a constant, which does not depend on θ. We will often use this proportional notation, i.e., if we explicitly calculate the posterior or predictive distribution.

Remark 3.1.23. *A general measure-theoretic version of Bayes formula can be found in Schervish (1995), which take-away message is: Given prior distribution Π and conditional weights or density p_θ one can construct a posterior distribution $\Pi \,|\, \{X = x\}$ and if Π has a density, the posterior distribution also has a density.*

3.2 Introductory example in Bayesian inference

In this section we choose the easy example of a coin toss with random success probability to introduce to Bayesian inference and motivate the main theoretic results about posterior prediction. Therefore, we start in Section 3.2.1 with the prior choice and demonstrate the advantage of a conjugate prior distribution. Furthermore, we illustrate the posterior update procedure and how to interpret this with probability statements. In Section 3.2.2 we explicitly check the posterior consistency for the chosen prior distribution. Afterwards in Section 3.2.3 we compare the Bayesian approach to classical statistic and highlight the advantage but also difficulty of Bayesian inference in case of a small sample size. If the sample size is large, we show that both approaches yield almost the same result. However, if we choose a bad prior distribution, which has no mass on the true value of θ, the posterior distribution is not consistent and does not equal the maximum likelihood estimator asymptotically, see Section 3.2.4.

3.2.1 Prior choice and posterior distribution

Imagine you are a gambler and you have a lot of biased coins. Assume you know that half of your coins are very unfair with a success probability ranges between 0.01 to 0.2 and the other half is unfair with a success probability ranges between 0.2 to 0.4. Furthermore, you know that in average the success probability is around 0.2. Unfortunately you put all your coins in one basket and can not distinguish the two types of coins. However, there is a person waiting to gamble against you, whereby you absolutely need to know if you choose a very unfair or an unfair coin. As the other person is very impatient, you only have time to do 10 coin tosses. How would you proceed?

Let us first mention that we are in a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ with $P_\theta = Ber(\theta)$, i.e., given the success probability $\{\Theta = \theta\}$ each coin toss X_i is independent and

$$X_i = \begin{cases} 1, & \text{with } \mathbb{P}(X_i = 1) = \theta, \\ 0, & \text{with } \mathbb{P}(X_i = 0) = 1 - \theta, \end{cases} \qquad \forall\, i = 1, \dots, 10.$$

For the number of successes $Y = \sum_{i=1}^{10} X_i$ it holds $Y \,|\, \{\Theta = \theta\} \sim Binom(10, \theta)$. According to our prior information we should choose a prior distribution, which assigns substantial mass to the interval $[0.01, 0.4]$, has an expectation of nearly 0.2 and assigns half of the mass to the area left of $\theta = 0.2$. However, there are infinitely many probability distributions that satisfy these conditions and according to the prior information there is no reasonable issue, why to choose a special one.

But we have a very impatient opponent, thus we choose a prior information, which has the required features and is very easy to handle. Here in case of a binomial sample distribution a Beta prior distribution is very suitable (for more details see Section 3.4):

As the Beta distribution has probability density function

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}, \quad \theta \in [0, 1],$$

where $\Gamma(\cdot)$ donates the Gamma function (Definition B.1), we obtain with Bayes formula for the posterior distribution:

$$\begin{aligned} \pi(\theta \,|\, Y = y) &= \frac{\pi(\theta) p_\theta(y)}{p(y)} \sim \pi(\theta) p_\theta(y) \\ &\sim \theta^{a-1} (1-\theta)^{b-1} \theta^y (1 - \theta^{10-y}) \\ &\sim \theta^{a+y-1} (1-\theta) b + 10 - y - 1 \sim Beta(a + y, b + 10 - y), \end{aligned}$$

thus again a Beta distribution. The pleasant part of so-called conjugate prior distribution is, that we only need to change the parameters of the prior distribution to involve the information, we obtained by observing a realization $Y = y$.

To involve our prior information, we notice that if $\Theta \sim Beta(a, b)$ then $\mathbb{E}[\Theta] = a/(a + b)$, cf. Remark B.3. If we choose $a = 4$ and $b = 15$, the expectation is around 0.2 and it holds $\mathbb{P}(\Theta \leq 0.2) \approx 0.5$, cf. Figure 3.1 A.

Assume we observe no success, i.e., $Y = 0$. Then the posterior distribution of Θ is

$$\Theta \,|\, \{Y = 0\} \sim Beta(4, 25),$$

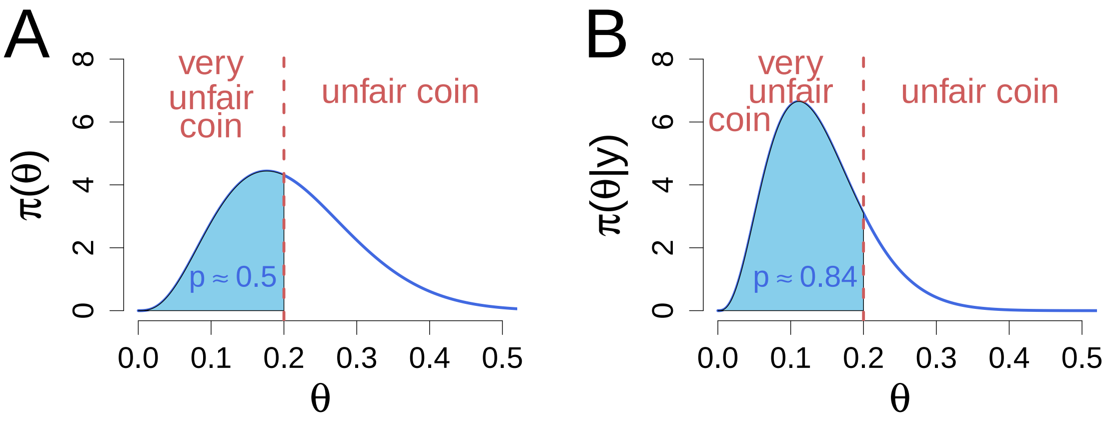


Figure 3.1: A. We choose a $Beta(4, 15)$ prior distribution, which has half of the mass left to $\theta = 0.2$ and expectation of almost 0.2. B. After observing 10 coin tosses with no success, our posterior belief changes into a $Beta(4, 25)$ distribution. Now 84% of the mass is left to $\theta = 0.2$.

which is shown in Figure 3.1 B. The posterior distribution is further to the left and more peaked. The shift to the left is due to the observation $Y = 0$, as this provides an evidence for a lower value of Θ. It is more peaked, because we have additional information due to the observed coin tosses. The prior expectation of Θ was about 0.2, the posterior expectation is about 0.14.

Determining the posterior distribution yields a learning model, which combines the prior information of Θ with the information we obtain from the sample. If we start with a prior belief of a $\mathcal{B}eta(4, 15)$ distribution, we have now faith in a $Beta(4, 25)$ distribution. With that we are able to determine magnitudes like $\mathbb{P}(\Theta \leq 0.2 \,|\, Y = 0) \approx 0.84$, which convinces us that we have drawn a very unfair coin.

3.2.2 Posterior consistency

In the coin toss example with a beta prior distribution Π we can easily check the consistency of the posterior distribution $\Pi \,|\, X_1, \dots, X_k$, which roughly spoken means: If θ_{true} is the true value of the parameter, then the posterior distribution $\Pi \,|\, X_1, \dots, X_k$ will degenerate at θ_{true} with probability 1 for $k \to \infty$. In Section 3.3.2 we will see that the consistency holds in general as long as the prior distribution assigns some weight to θ_{true}, what a Beta distribution fulfills for all $\theta \in (0, 1)$.

Let us define $Y_k := X_1 + \cdots, X_k$. Then we know for the posterior distribution, that $\Theta \,|\, Y_k \sim Beta(a + Y_k, b + 10 - Y_k,)$, so we get

$$\mathbb{E}[\Theta \,|\, Y_k] = \frac{Y_k + a}{k + a + b} \quad \text{and} \quad \mathbb{V}ar[\Theta \,|\, Y_k] = \frac{(Y_k + a)(k - Y_k + b)}{(k + a + b)^2 (k + a + b + 1)}.$$

Defining $\bar{Y}_k := \frac{1}{k} Y_k$ and rewriting the formula of the expectation and the variance yields

$$\mathbb{E}[\Theta \,|\, Y_k] = \frac{k\bar{Y}_k + a}{k + a + b} \quad \text{and} \quad \mathbb{V}ar[\Theta \,|\, Y_k] = \frac{k^2(\bar{Y}_k + a/k)(1 - \bar{Y}_k + b/k)}{n^3(1 + (a + b)/k)^2(1 + (a + b + 1)/k)}.$$

Due to the law of large number $\bar{Y}_k \overset{k\to\infty}{\longrightarrow} \theta_{true}$ a.s. and hence

$$\mathbb{E}[\Theta \,|\, Y_k] \overset{k\to\infty}{\longrightarrow} \theta_{true} \text{ a.s.} \qquad \text{and} \qquad \mathbb{V}ar[\Theta \,|\, Y_k] \overset{k\to\infty}{\longrightarrow} 0 \text{ a.s.},$$

so the posterior distribution collapses at θ_0.

3.2.3 Comparison to classical statistic

In the classical statistic we would consider the likelihood function

$$f_{y_k}(\theta) = \binom{k}{y_k} \theta^{y_k} (1-\theta)^{k-y_k}.$$

Following the maximum likelihood approach, we calculate

$$\ell(\theta) := -\log(f_{y_k}(\theta)) = \log\left(\binom{k}{y_k}\right) + y_k \log(\theta) + (k - y_k)\log(1-\theta).$$

By differentiating and setting to zero, we can determine the value of θ, which maximizes the likelihood:

$$\begin{aligned} \frac{\partial \ell(\theta)}{\partial \theta} &= \frac{y_k}{\theta} - \frac{k-y_k}{1-\theta} \overset{!}{=} 0 \\ (1-\theta)y_n &= (k-y_k)\theta \\ \Rightarrow \hat{\theta}_{ML} &= \frac{y_k}{k} =: \bar{y}_k. \end{aligned}$$

Back to our coin toss example, where we observe no success in 10 rounds, we would estimate $\hat{\theta}_{ML} = 0/10 = 0$, which even does not fit to our prior information (we know, there is at least a success probability of 0.01). However, a point-wise estimator is quite implausible for the small sample size, but even the use of a standard confidence interval by the normal approximation does not yield a usefulness description of $\hat{\theta}$, as

$$\bar{y}_k \pm 1.96\sqrt{\bar{y}_k(1-\bar{y}_k)/k} = 0.$$

However, using the ClopperPearson interval would be a way out here.

Maximum a posterior (MAP) estimation

Following Bayes inference we get a posterior distribution, which reflects our opinion and combines the prior information and the observed sample. If we are interested in just one 'best' value of θ, instead of finding θ, that maximizes the likelihood function $f_{y_k}(\cdot)$, we could search for θ, which maximizes the posterior density $\pi(\cdot \,|\, Y_k = y_k)$. So we would choose

$$\hat{\theta}_{MAP} \in arg\max_\theta \pi(\theta \,|\, Y_k = y_k) = arg\max_\theta \frac{p_\theta(y_k)\pi(\theta)}{\int_0^1 p_{\tilde{\theta}}(y_k)\pi(\tilde{\theta})\, d\tilde{\theta}}.$$

As the integral does not depend on the value of θ, we can simplify

$$\hat{\theta}_{MAP} \in arg\max_\theta p_\theta(y_k)\pi(\theta) = arg\max_\theta (\log(p_\theta(y_k)) + \log(\pi(\theta))). \tag{3.1}$$

Thus if the prior distribution $\pi(\cdot)$ is uniform on $[0,1]$, the MAP estimator $\hat{\theta}_{MAP}$ is identical to the maximum likelihood estimator $\hat{\theta}_{ML}$.
Based on Equation (3.1) we can obviously see, that the chosen prior distribution $\pi(\cdot)$ has a strong impact on the MAP estimator $\hat{\theta}_{MAP}$, especially in case of a small sample size. However, as long as we start with a prior distribution, that assigns non-zero probability to the true value θ_{true}, the MAP estimator is consistent and the maximum likelihood and MAP-estimator are asymptotically identical.

In our example we choose a prior belief of $\Theta \sim Beta(a,b)$, which yields a nice formula for the MAP estimator: Using Equation (3.1) and plugging in the likelihood of our sampling model and the density of the Beta distribution, we get

$$\hat{\theta}_{MAP} \in arg \max_{\theta} \left(y_k \log(\theta) + (k-y_k)\log(1-\theta) + (a-1)\log(\theta) + (b-1)\log(1-\theta)\right).$$

Differentiating and setting to zero yields

$$\begin{aligned}
\frac{y_k}{\theta} - \frac{k-y}{1-\theta} + \frac{a-1}{\theta} - \frac{b-1}{1-\theta} &\overset{!}{=} 0 \\
\frac{y_k+a-1}{\theta} &= \frac{k-y_k+b-1}{1-\theta} \\
(y_k+a-1)(1-\theta) &= (k-y_k+b-1)\theta \\
\theta(k+a-1+b-1) &= y_k+a-1 \\
\Rightarrow \hat{\theta}_{MAP} &= \frac{y_k+a-1}{(k+b-1)+a-1}.
\end{aligned}$$

Comparing this to the maximum likelihood estimator $\hat{\theta}_{ML} = y_k/k$ shows: The MAP estimator is equal to a maximum likelihood estimator of a coin toss with additional $a-1+b-1$ throws, whereby we observe $a-1$ successes and $b-1$ failures.
Furthermore, we can recognize directly, the MAP estimator equals the maximum likelihood estimator asymptotically, as an additional finite number a and b do not destroy the convergence. However, in case of a small sample size a thoughtful choice of a and b helps to generate a plausible estimator of Θ.

Posterior expectation
The involvement of Bayesian inference and maximum likelihood estimation can be also found, rewriting the posterior expectation, starting with a prior distribution $\Theta \sim Beta(a,b)$:

$$\begin{aligned}
\mathbb{E}[\Theta \,|\, Y_k = y_k] &= \frac{a+y_k}{a+b+k} \\
&= \frac{n}{a+b+k}\frac{y_k}{k} + \frac{a+b}{a+b+k}\frac{a}{a+b} \\
&= \frac{k}{\omega+k}\hat{\theta}_{ML} + \frac{\omega}{\omega+k}\theta_0 \qquad \overset{k\to\infty}{\longrightarrow} \hat{\theta}_{ML},
\end{aligned}$$

where $\omega := a+b$ and $\theta_0 := \mathbb{E}[\Theta]$ the prior expectation. Hereby ω can be interpreted as confidence in the prior distribution. If we have a higher confidence in the prior distribution and the prior is well-chosen, we will get even for small n a reasonable and stable estimation of Θ. Otherwise, if we have a high confidence in the prior distribution, but the prior information is bad, we need a large sample size to counter the bias and the maximum likelihood estimating would be more appropriate.

In Section 3.4.4 we will see, that in case of an exponential family distribution (what a Binomial distribution is), there exists a standard conjugate prior distribution form, and if we choose this prior, we always have posterior linearity in the sufficient statistic of the canonical parameter, see also Example 3.4.6 and 3.4.15. Under some weak regulatory conditions the inversion is also true, that if we have the posterior linearity in the sufficient statistic, the prior distribution of the canonical parameter must be of the standard conjugate prior form, cf. Theorem 3.4.24.

3.2.4 Bad prior choice

We have already mentioned, that it is crucial for posterior consistency to choose a prior which assigns some mass to the true value of θ. Here we will show on basis of the coin toss example, what happens if we choose a bad prior distribution: Assume our prior information is bad and the true success probability of the chosen coin is $\theta_{true} = 0.5$. If we had not chosen a Beta prior, but a uniform distribution, i.e $\pi(\theta) = 2.5$ for $\theta \in [0, 0.4]$, we need to calculate the posterior distribution numerically by

$$\pi(\theta \,|\, Y_k = y_k) = \frac{\theta^{y_k}(1-\theta)^{k-y_k} \cdot 2.5}{\int_0^{0.4} \tilde{\theta}^{y_k}(1-\tilde{\theta})^{k-y_k} \cdot 2.5 \, d\tilde{\theta}}, \quad \text{for } \theta \in [0, 0.4].$$

In Figure 3.2 the posterior distributions can be found for $n = 10, 100, 200$ and realizations $y_n = 5, 48, 96$. Since the prior distributions only assigns weight to values in $[0, 0.4]$, the

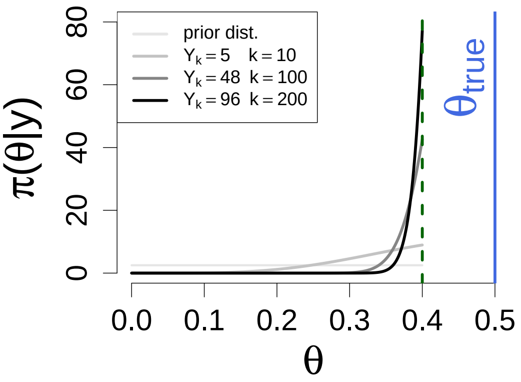


Figure 3.2: Consider a coin toss $Y_k \sim Binom(n, \Theta)$ and choose a uniform prior distribution $\pi(\theta) = 2.5$ for $\theta \in [0, 0.4]$ (light gray). Assume we have bad prior information and the true parameter of the coin toss is $\theta_{true} = 0.5$ (blue). The corresponding posterior distributions for $k = 10, 100, 200$ and $y_k = 5, 48, 96$ are shown in darker and darker tones.

posterior distributions can not left this area. However, the posterior distributions shifts on and on to the right border 0.4. In Section 3.3.2 we will see, that the posterior distribution asymptotically concentrates on the value, for which the likelihood function is closest to the true generating distribution in sense of the KullbackLeibler divergence. Here the sampling model is specified correctly, so the posterior distribution concentrates on the value, which is closest to the true value θ_{true}.

3.3 How to choose a proper prior distribution?

A proper prior distribution should at least asymptotically yield a posterior distribution that degenerates at the true parameter value, which is meant by posterior consistency. But what if the Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ is not correct and the observed samples $X_1, X_2 \ldots$ arise from a distribution Q with $Q \neq P_\theta$ for all θ? Which value of θ will we observe? Or, if the sample distribution is chosen correctly, which technical requirements do we need for the prior distribution to obtain posterior consistency?
To answer these questions asymptotically we use the Kullback-Leibler divergence, which is a measure of the difference between two probability distributions, see Section 3.3.1.
In Section 3.3.2 we show that we will end in the value θ^*, for which the wrong sample model is 'closest' to the true generating distribution Q in sense of the Kullback-Leibler divergence. If the sample distribution is correct, than we will have posterior consistency as long as the prior distribution assigns some weight to the true value θ_{true}.
Under some regulatory assumptions a stronger result can be proved, cf. Section 3.3.3: The posterior distribution is asymptotic normally distributed with mean the maximum likelihood estimator and variance the inverse Fisher-information declining with the sample size. As the proof of the general version is quite technical, we motivate the result by two examples, whereby the first is fairly obvious.

3.3.1 Some properties of the Kullback-Leibler divergence

Definition 3.3.1 (Absolute continuous). *Let Q and P be two probability distributions with densities $q(\cdot)$ and $p(\cdot)$ or discrete weights $q(\cdot)$ and $p(\cdot)$. Q is absolute continuous with respect to P, if $p(x) = 0$ implies $q(x) = 0$ $\forall x \in \mathbb{R}$.*

Definition 3.3.2 (Kullback-Leibler divergence). *Let Q and P be two probability distributions and Q be absolute continuous with respect to P. Furthermore let $X \sim Q$. For discrete probability distributions with weights $q(\cdot)$ and $p(\cdot)$, the Kullback-Leibler divergence is defined as*

$$D_{KL}(Q||P) := \mathbb{E}_Q \left[\log \frac{q(X)}{p(X)} \right] = \sum_x q(x) \log \frac{q(x)}{p(x)}.$$

For continuous distributions with densities $q(\cdot)$ and $p(\cdot)$, the Kullback-Leibler divergence is defined as

$$D_{KL}(Q||P) := \mathbb{E}_Q \left[\log \frac{q(X)}{p(X)} \right] = \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)} \, dx.$$

Remark 3.3.3. *Sometimes the Kullback-Leibler divergence is called a distance, but in general the requested property of symmetry does not hold, as in general $D_{KL}(Q||P) \neq D_{KL}(P||Q)$.*

As average of the logarithmic difference between probability distributions Q and P, where the average is taken with respect to Q, the Kullback-Leibler divergence is non-negative. To proof this statement we use Jensen's inequality:

Lemma 3.3.4 (Jensen inequality). *For every convex function $K : \mathbb{R} \to \mathbb{R}$ and every random variable X with finite expectation it holds*

$$\mathbb{E}[K(X)] \geq K\left(\mathbb{E}[X]\right).$$

Proof. See Kersting and Wakolbinger (2010). A convex function $K(\cdot)$ has the property that for every $d \in \mathbb{R}$ there exists a linear function $g(x) = K(d) + c(x - d)$ with

$$g(a) \leq K(a) \text{ for all } a \in \mathbb{R} \quad \text{and} \quad g(d) = K(d).$$

Replacing a by the random variable X and taking the expectation yields

$$K(d) + c(\mathbb{E}[X] - d) \leq \mathbb{E}[K(X)],$$

due to the linearity and monotony of the expectation. Choosing $d = \mathbb{E}[X]$ gives the statement. □

Remark 3.3.5. *If $K(\cdot)$ is a strictly convex function, the equality in Lemma 3.3.4 holds if and only if X is constant. In that case $\mathbb{E}[X] = X$ and $\mathbb{E}[K(X)] = K(X)$, which yields equality for X constant.*

Lemma 3.3.6. *Let Q and P be two probability distributions with densities $q(\cdot)$ and $p(\cdot)$ and let Q be absolute continuous with respect to P. The Kullback-Leibler divergence is non-negative, i.e.,*

$$D_{KL}(Q||P) \geq 0,$$

with $D_{KL}(Q||P) = 0$ if and only if $Q = P$ almost everywhere.

Proof. Let Y be a random variable with distribution Q. We define the random variable $X := \frac{p(Y)}{q(Y)}$. Furthermore, we define $K(\cdot) := -\log(\cdot)$, a strictly convex function. Applying Jensen's inequality (Lemma 3.3.4) yields

$$\begin{aligned} \mathbb{E}_Q[K(X)] \geq& K(\mathbb{E}[X]) \\ \implies \quad -\int_{-\infty}^{\infty} q(y) \log \frac{p(y)}{q(y)}\, dy \geq& -\log\left(\int_{-\infty}^{\infty} q(y) \frac{p(y)}{q(y)}\, dy\right) \\ \implies \quad \int_{-\infty}^{\infty} q(y) \log \frac{q(y)}{p(y)}\, dy \geq& -\log\left(\int_{-\infty}^{\infty} p(y)\, dy\right) \\ \implies \quad \int_{-\infty}^{\infty} q(y) \log \frac{q(y)}{p(y)}\, dy \geq& -\log(1) = 0. \end{aligned}$$

According to Remark 3.3.5 the equality holds if and only if X is constant, so $p(y) = q(y)$ almost everywhere. □

Remark 3.3.7. *Lemma 3.3.6 holds obviously also for discrete probability distribution Q and P.*

Remark 3.3.8. *The Kullback-Leibler divergence is invariant under parameter transformation: Therefore we consider a transformation from variable X to variable $t(X) = Y$ and rewrite the Kullback-Leibler divergence*

$$\begin{aligned}
D_{KL}(Q_x||P_x) &= \int_{x_a}^{x_b} q(x) \log \frac{q(x)}{p(x)}\, dx \\
&= \int_{t(x_a)}^{t(x_b)} q(y) \log \frac{q(y)dy/dx}{p(y)dy/dx}\, dy = \int_{t(x_a)}^{t(x_b)} q(y) \log \frac{q(y)}{p(y)}\, dy = D_{KL}(Q_y||P_y.)
\end{aligned}$$

Remark 3.3.9. *The Kullback-Leibler divergence is additive for independent distributions: If Q_1, Q_2 are independent distributions with joint distribution $Q(x,y) = Q_1(x) \cdot Q_2(y)$ and P, P_1, P_2 likewise, then*

$$\begin{aligned}
D_{KL}(Q||P) &= \int_{-\infty}^{\infty} q(x,y) \log \frac{q(x,y)}{p(x,y)}\, d(x,y) \\
&= \int_{-\infty}^{\infty} q_1(x)q_2(y) \log \left(\frac{q_1(x)}{p_1(x)} \frac{q_2(y)}{p_1(y)} \right)\, d(x,y) \\
&= \int_{-\infty}^{\infty} q_1(x)q_2(y) \log \left(\frac{q_1(x)}{p_1(x)} \right)\, dy\, dx + \int_{-\infty}^{\infty} q_1(x)q_2(y) \log \left(\frac{q_2(y)}{p_2(y)} \right)\, dx\, dy \\
&= \int_{-\infty}^{\infty} q_1(x) \log \left(\frac{q_1(x)}{p_1(x)} \right)\, dx + \int_{-\infty}^{\infty} q_2(y) \log \left(\frac{q_2(y)}{p_2(y)} \right)\, dy \\
&= D_{KL}(Q_1||P_1) + D_{KL}(Q_2||P_2).
\end{aligned}$$

3.3.2 Consistency of the posterior distribution

Definition 3.3.10. *A sequence $X_1, X_2, \dots$ of random variables converges in probability to a random variable X, if for all $\epsilon > 0$*

$$\lim_{k\to\infty} \mathbb{P}(|X_k - X| \geq \epsilon) = 0.$$

We denote this convergence as $X_k \overset{\mathbb{P}}{\to} X$.

Definition 3.3.11 (Posterior consistency)**.** *Consider a sequence of independent random variables $X_1, \dots, X_k$ from a sample distribution $P_{\theta_{true}}$. Further, let Π be the prior distribution of a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$. The posterior distribution $\Pi \,|\, \{X_1 = x_1, \dots, X_k = x_k\}$ is named to be consistent, if for every $\epsilon > 0$*

$$\Pi\left(\{\theta : ||\theta - \theta_{true}|| > \epsilon\} \,|\, x_1, \dots, x_k\right) \overset{k\to\infty}{\longrightarrow} 0.$$

Considering posterior consistency raises two natural questions: First how does a bad specified sample distribution influence posterior consistency and second to what extend does a bad specified prior distribution threatens posterior consistency? To answer the first question we assume we observe an independent sample from an arbitrary distribution Q, but we believe in a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$. Then Theorem 3.3.12 tells us that we will end in the value θ^* for which the wrong sample model is 'closest' to the true generating distribution Q in sense of the Kullback-Leibler divergence. Therefore, its crucial that the prior distribution assigns some mass to the value θ^*, which leads us to the answer of the second question: If the sample distribution is chosen correctly and as long as the prior distribution assigns some weight to the true value θ_{true}, we will have posterior consistency.

Theorem 3.3.12 (Posterior consistency)**.** *Let $X_1, \dots, X_k$ be a sequence of independent samples from an arbitrary distribution Q. Erroneously we think the sample arises from a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$. We define*

$$\theta^* := arg \min_\theta D_{KL}(Q||P_\theta)$$

the parameter of the sample model $\mathcal{P}$, for which it is 'closest' to the true generating distribution Q and we assume there exist a unique minimizer. If for every $\epsilon > 0$

$$\Pi\left(\{\theta : \|\theta - \theta^*\| \leq \epsilon\}\right) > 0,$$

then

$$\Pi\left(\{\theta : \|\theta - \theta^*\| > \epsilon\} \,|\, X_{1:k} = x_{1:k}\right) \overset{k\to\infty}{\longrightarrow} 0.$$

Proof. We will prove Theorem 3.3.12 only for the case of a finite parameter space. An outline of proofs for discrete or continuous parameter space can be found in Gelman et al. (2013).

For any $\theta \neq \theta^*$ we show that $\pi(\theta \,|\, X_{1:k} = x_{1:k}) \to 0$ for a large sample set. Therefore we consider the logarithm of the posterior quotient of θ and θ^*:

$$\log \frac{\pi(\theta \,|\, X_{1:k} = x_{1:k})}{\pi(\theta^* \,|\, X_{1:k} = x_{1:k})} = \log \frac{p_\theta(x_1, \dots, x_k)\pi(\theta)}{p_{\theta^*}(x_1, \dots, x_k)\pi(\theta^*)} = \log\left(\frac{\pi(\theta)}{\pi(\theta^*)}\right) + \sum_{i=1}^k \log \frac{p_\theta(x_i)}{p_{\theta^*}(x_i)}. \tag{3.2}$$

By the law of large numbers we get

$$\frac{1}{k}\sum_{i=1}^k \log \frac{p_\theta(x_i)}{p_{\theta^*}(x_i)} \xrightarrow[k\to\infty]{\mathbb{P}} \mathbb{E}_Q\left[\log \frac{p_\theta(X)}{p_{\theta^*}(X)}\right]. \tag{3.3}$$

The expectation can be transformed to two Kullback-Leibler divergences

$$\begin{aligned}
\mathbb{E}_Q\left[\log \frac{p_\theta(X)}{p_{\theta^*}(X)}\right] &= \mathbb{E}_Q\left[\log \frac{p_\theta(X)q(X)}{p_{\theta^*}(X)q(X)}\right] \\
&= \mathbb{E}_Q\left[\log \frac{q(X)}{p_{\theta^*}(X)} - \log \frac{q(X)}{p_\theta(X)}\right] \\
&= D_{KL}(Q||P_{\theta^*}) - D_{KL}(Q||P_\theta) \\
&< 0, \qquad (3.4)
\end{aligned}$$

as according to Lemma 3.3.6 the Kullback-Leibler divergence is non-negative and we assumed that θ^* is a unique minimizer.
Due to Equation (3.3) and 3.4 we obtain

$$\sum_{i=1}^k \log \frac{p_\theta(x_i)}{p_{\theta^*}(x_i)} \xrightarrow[k\to\infty]{\mathbb{P}} k \cdot \mathbb{E}_Q\left[\log \frac{p_\theta(X)}{p_{\theta^*}(X)}\right] = -\infty.$$

As the prior distribution does not depend on k and we assumed $\pi(\theta^*) > 0$, Equation (3.2) can be determined

$$\log \frac{\pi(\theta \,|\, X_{1:k} = x_{1:k})}{\pi(\theta^* \,|\, X_{1:k} = x_{1:k})} = \log\left(\frac{\pi(\theta)}{\pi(\theta^*)}\right) + \sum_{i=1}^k \log \frac{p_\theta(x_i)}{p_{\theta^*}(x_i)} \xrightarrow[k\to\infty]{\mathbb{P}} -\infty$$

and

$$\log \frac{\pi(\theta \,|\, X_{1:k} = x_{1:k})}{\pi(\theta^* \,|\, X_{1:k} = x_{1:k})} \xrightarrow[k\to\infty]{\mathbb{P}} -\infty \quad \text{implies} \quad \frac{\pi(\theta \,|\, X_{1:k} = x_{1:k})}{\pi(\theta^* \,|\, X_{1:k} = x_{1:k})} \xrightarrow[k\to\infty]{\mathbb{P}} 0,$$

which implies $\pi(\theta \,|\, X_{1:k} = x_{1:k}) \xrightarrow[k\to\infty]{\mathbb{P}} 0$ for every $\theta \neq \theta^*$. □

Remark 3.3.13. *In Theorem 3.3.12 we assumed there exists a unique minimizer θ^* of the Kullback-Leibler divergence to the true generating distribution Q. Assume we choose the sampling model $\mathcal{P}$ correctly, i.e there exists a θ_{true} such that $q(x) = p_{\theta_{true}}(x)$ for all $x \in \mathbb{R}$. Then we know from Lemma 3.3.6, that $\theta^* = \theta_{true}$, as θ_{true} minimizes the Kullback-Leibler divergence ($D_{KL}(Q||P_{\theta^*}) = 0$). If there exists a θ_1 with $D_{KL}(Q||P_{\theta_1}) = 0$, then $P_{\theta_{true}} = P_{\theta_1}$ almost everywhere, thus θ_{true} is a unique minimizer cf. Lemma 3.3.6.*

Even with the knowledge that as long as the prior distribution assigns some mass to the true value θ_{true}, we will end at the true parameter, this can be a difficult task: If we do not have specific prior information, we must assign some mass to every plausible value of θ. To get around we could choose a very broad prior distribution. Then we will very likely end in the true value of θ, but we should recall, that Theorem 3.3.12 is only an asymptotic statement. But the strength of Bayesian inference is just to integrate prior information in the parameter estimation to get qualified estimations, even for small sample sizes. With a broad prior distribution we can not benefit from Bayesian inference in that way. However one should bear in mind that in case of a small sample sizes, poor choices of the prior distribution or the sample model we will probably get poor results and we must be cautious.

3.3.3 Asymptotic normality of the posterior distribution

In the following we assume that the Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ is chosen correctly, thus there exists a value θ_{true} of the parameter space, such that $P_{\theta_{true}}$ is the true sample distribution. Under some regulatory assumptions a stronger result than Theorem 3.3.12 can be proved:
If $\hat{\theta}_{ML}^{(k)}$ is the maximum likelihood estimator based on the sample $X_1, \ldots, X_k$ with distribution $P_{\theta_{true}}$, then the posterior distribution is asymptotically equal to a normal distribution with mean $\hat{\theta}_{ML}^{(k)}$ and variance $\mathcal{I}^{-1}(\hat{\theta}_{ML}^{(k)})/k$, where $\mathcal{I}(\cdot)$ denotes the Fisher-information.

This result dates back to Laplace, who proved the posterior normality for the special case of a constant prior distribution. Richard von Mises proved the posterior normality, if the prior distribution has a continuous and restricted density. Similar considerations were made by the Russian analyst S. N. Bernstein. Finally, L. Le Cam extended the proof under more general assumptions. The result is known as Bernstein-von Mises theorem or Bayesian central limit theorem.

Example 3.3.14. *Consider a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ with Π a standard normal distribution, i.e.,*

$$\Theta \sim \mathcal{N}(0, 1),$$

and $\mathcal{P}$ a family of normal distribution with variance 1*, i.e., given* $\{\Theta = \theta\}$

$$X_i \sim \mathcal{N}(\theta, 1) \quad i = 1, 2, \ldots.$$

Later in Lemma 4.3.1 we will see, that the posterior distribution is given bv

$$\Theta \,|\, \{X_{1:k} = x_{1:k}\} \sim \mathcal{N}\left(\frac{\sum_{i=1}^{k} x_i}{k+1}, \frac{1}{k+1} \right).$$

Calculating the Fisher-information for $X \sim \mathcal{N}(\theta, \sigma^2)$ yields

$$\begin{aligned} \mathcal{I}(\theta) :=& \mathbb{E}\left[-\frac{\partial^2}{\partial \theta} \log(p_\theta(x)) \right] \\ =& \mathbb{E}\left[-\frac{\partial}{\partial \theta} \frac{x-\theta}{\sigma^2} \right] \\ =& \mathbb{E}\left[\frac{1}{\sigma^2} \right] = \frac{1}{\sigma^2} \end{aligned}$$

and as the maximum likelihood estimator of θ is equal to the mean $\bar{x} := 1/k \sum x_i$, according to the Bernstein-von Mises theorem, see Theorem 3.3.16, the posterior distribution should be asymptotically

$$\Theta \,|\, \{X_{1:k} = x_{1:k}\} \overset{a}{\sim} \mathcal{N}\left(\frac{\sum_{i=1}^{k} x_i}{k}, \frac{1}{k} \right),$$

which is true according to Lemma 4.3.1.

Example 3.3.15. *Returning to the introductory example we consider a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ with $\Theta \sim Beta(a, b)$ and given $\{\Theta = \theta\}$ we choose $Y_k \sim Binom(k, \theta)$. We have already seen, that the posterior distribution is $\Theta \,|\, \{Y_k = y_k\} \sim Beta(a + y_k, b + k - y_k)$. Let $\hat{\theta}_{ML}^{(k)} = y_k/k$ denote the maximum likelihood estimator of θ.*
For large k we can approximate the posterior distribution by

$$\Theta \,|\, \{Y_k = y_k\} \sim Beta(k\hat{\theta}_{ML}^{(k)}, k(1 - \hat{\theta}_{ML}^{(k)})).$$

Using the connecting to the Gamma distribution (see Remark B.8), that if $X \sim Gamma(p_1, m)$ and $Z \sim Gamma(p_2, m)$ and both independent, than

$$\frac{X}{X+Z} \sim Beta(p_1, p_2)$$

and the property (see Remark B.7), that if $E_1, \ldots, E_k \sim Exp(1)$ and independent, than

$$\sum_{i=1}^{k} E_i \sim Gamma(k, 1),$$

yield

$$\Theta \,|\, \{Y_k = y_k\} \sim \frac{X}{X+Z} \sim \frac{\sum_{i=1}^{k\hat{\theta}_{ML}^{(k)}} E_i}{\sum_{i=1}^{k\hat{\theta}_{ML}^{(k)}} E_i + \sum_{i=k\hat{\theta}_{ML}^{(k)}+1}^{k+1} E_i},$$

where $X \sim Gamma(k\hat{\theta}_{ML}^{(k)}, 1)$, $Z \sim Gamma(k(1-\hat{\theta}_{ML}^{(k)}), 1)$ and $E_i \sim Exp(1)$, $i = 1, \ldots, k+1$ and independent. Some algebra gives us

$$\begin{aligned}
&\sqrt{k}\left(\frac{\sum_{i=1}^{k\hat{\theta}_{ML}^{(k)}} E_i}{\sum_{i=1}^{k\hat{\theta}_{ML}^{(k)}} E_i + \sum_{i=k\hat{\theta}_{ML}^{(k)}}^{k+1} E_i} - \hat{\theta}_{ML}^{(k)}\right)\\
&= \frac{\frac{1}{\sqrt{k}}\left((1-\hat{\theta}_{ML}^{(k)})\left(\sum_{i=1}^{k\hat{\theta}_{ML}^{(k)}} E_i - k\hat{\theta}_{ML}^{(k)}\right) - \hat{\theta}_{ML}^{(k)}\left(\sum_{i=k\hat{\theta}_{ML}^{(k)}+1}^{k+1} E_i - (k - k\hat{\theta}_{ML}^{(k)}+1)\right)\right)}{\sum_{i=1}^{k+1} E_i/k}\\
&+ \frac{\frac{1}{\sqrt{k}}\left((1-\hat{\theta}_{ML}^{(k)})k\hat{\theta}_{ML}^{(k)} - \hat{\theta}_{ML}^{(k)}(k - k\hat{\theta}_{ML}^{(k)}+1)\right)}{\sum_{i=1}^{k+1} E_i/k}
\end{aligned}$$

and we observe for the third part

$$\frac{(1-\hat{\theta}_{ML}^{(k)})k\hat{\theta}_{ML}^{(k)} - \hat{\theta}_{ML}^{(k)}(k - k\hat{\theta}_{ML}^{(k)}+1)}{\sqrt{k}} = \frac{\hat{\theta}_{ML}^{(k)}}{\sqrt{k}} \overset{k\to\infty}{\longrightarrow} 0.$$

Since $\mathbb{E}[E_i] = 1$ and $\mathbb{V}ar[E_i] = 1$ and the E_i's are i.i.d., it follows from the central limit theorem that

$$\frac{\sum_{i=1}^{k\hat{\theta}_{ML}^{(k)}} E_i - k\hat{\theta}_{ML}^{(k)}}{\sqrt{k\hat{\theta}_{ML}^{(k)}}} \overset{a}{\sim} \mathcal{N}(0,1), \quad \textit{thus} \quad \frac{\sum_{i=1}^{k\hat{\theta}_{ML}^{(k)}} E_i - k\hat{\theta}_{ML}^{(k)}}{\sqrt{k}} \overset{a}{\sim} \mathcal{N}(0, \hat{\theta}_{ML}^{(k)})$$

and analogously that

$$\frac{\sum_{i=k\hat{\theta}_{ML}^{(k)}+1}^{k+1} E_i - (k - k\hat{\theta}_{ML}^{(k)}+1)}{\sqrt{k}} \overset{a}{\sim} \mathcal{N}(0, 1-\hat{\theta}_{ML}^{(k)}).$$

As $\sum_{i=1}^{k\hat{\theta}_{ML}^{(k)}} E_i$ and $\sum_{i=1}^{k\hat{\theta}_{ML}^{(k)}} E_i$ are independent for all k and

$$\left(1-\hat{\theta}_{ML}^{(k)}\right)^2 \hat{\theta}_{ML}^{(k)} + (\hat{\theta}_{ML}^{(k)})^2\left(1-\hat{\theta}_{ML}^{(k)}\right) = \hat{\theta}_{ML}^{(k)}\left(1-\hat{\theta}_{ML}^{(k)}\right),$$

we get

$$\left(1-\hat{\theta}_{ML}^{(k)}\right)\frac{\sum_{i=1}^{k\hat{\theta}_{ML}^{(k)}} E_i - k\hat{\theta}_{ML}^{(k)}}{\sqrt{k}} - \hat{\theta}_{ML}^{(k)}\frac{\sum_{i=k\hat{\theta}_{ML}^{(k)}+1}^{k+1} E_i - (k - k\hat{\theta}_{ML}^{(k)}+1)}{\sqrt{k}} \overset{a}{\sim} \mathcal{N}\left(0, \hat{\theta}_{ML}^{(k)}\left(1-\hat{\theta}_{ML}^{(k)}\right)\right).$$

With law of large numbers we obtain for the denominator

$$\sum_{i=1}^{k+1} E_i \xrightarrow{\mathbb{P}} 1.$$

Applying Slutsky's theorem, which states for sequences of random variables X_k and Z_k, where X_k converges in distribution to a random variable X and Z_k converges in probability to a constant c, then $Z_k X_k \xrightarrow{\mathbb{P}} cX$, we overall obtain

$$\Theta \,|\, \{Y_k = y_k\} \xrightarrow{\mathbb{P}} \mathcal{N}\left(\hat{\theta}_{ML}^{(k)}, \frac{1}{k}\hat{\theta}_{ML}^{(k)}\left(1-\hat{\theta}_{ML}^{(k)}\right)\right). \tag{3.5}$$

We can verify this expression with Theorem 3.3.16 by calculating the Fisher-information

$$\begin{aligned}\mathcal{I}(\theta) &= \mathbb{E}\left[-\frac{\partial^2}{\partial\theta}(y\log(\theta)+(1-y)\log(1-\theta))\right]\\ &= \mathbb{E}\left[-\frac{\partial}{\partial\theta}\left(\frac{y}{\theta}+\frac{1-y}{1-\theta}\right)\right]\\ &= \mathbb{E}\left[\frac{y}{\theta^2}+\frac{1-y}{(1-\theta)^2}\right]=\frac{1}{\theta}+\frac{1}{1-\theta}=\frac{1}{\theta(1-\theta)},\end{aligned}$$

which confirms Equation (3.5)*.*

To proof Theorem 3.3.16 some regulatory conditions are needed. Basically these assumptions are identically to those, which are needed to prove the asymptotically normality of the maximum likelihood estimator. For the Bernstein-von Mises Theorem we require additional, that the prior density $\pi(\cdot)$ is continuous and θ_{true} is not on the boundary. Furthermore, we need some mass at the true value θ_{true}, which is already needed for posterior consistency. The regulatory conditions can be found in Ghosh and Ramamoorthi (2003).

Theorem 3.3.16 (Bernstein-von Mises Theorem)**.** *Consider a Bayesian model $\mathcal{B}(\Pi,\mathcal{P})$ and let $X_1,\dots,X_k$ be an i.i.d. sample with distribution $P_{\theta_{true}}$. Further, let $\hat{\theta}_{ML}^{(k)}$ denote the maximum likelihood estimator. Then under some regulatory conditions, see (Ghosh and Ramamoorthi, 2003), the posterior distribution of Θ is asymptotically normally distributed, i.e.,*

$$\Theta\,|\,\{Y_k=y_k\}\xrightarrow{\mathbb{P}}\mathcal{N}\left(\hat{\theta}_{ML}^{(k)},\frac{1}{k}\mathcal{I}^{-1}\left(\hat{\theta}_{ML}^{(k)}\right)\right),$$

where $\mathcal{I}(\cdot)^{-1}$ denotes the inverse Fisher-information, i.e.,

$$\mathcal{I}(\theta)=\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log p_\theta(x)\right].$$

Proof. A proof sketch can be found in Gelman et al. (2013), a more technical version in Ghosh and Ramamoorthi (2003). □

3.4 Conjugate priors and the general case of an exponential family distribution

To use Bayesian inference to search for change points, cf. Section 4.2, we need access to the predictive distribution $p(x_{k+1}\,|\,X_{1:k}=x_{1:k})$. In general, we need to evaluate the integral

$$\begin{aligned}p(x_{k+1}\,|\,X_{1:k}=x_{1:k}) &= \int p_\theta(x_{k+1})\cdot\pi(\theta\,|\,X_{1:k}=x_{1:k})d\theta\\ &= \int p_\theta(x_{k+1})\cdot\frac{\pi_\theta(\theta)p_\theta(x_1,\dots,x_k)}{p(x_1,\dots,x_k)}d\theta,\end{aligned}$$

which can be very difficult or even impossible. In that case Markov Chain Monte Carlo steps can be used, resulting in high computational cost.

If we only consider sampling models belonging to the exponential family distribution, we can choose a prior distribution $\pi_\theta(\cdot)$ on θ, such that the posterior distribution $\pi_\theta(\cdot\,|\,X_{1:k}=x_{1:k})$

is in the same distribution family as the prior. Such prior is called conjugate. In case of an exponential family sampling model even the predictive distribution $p(x_{k+1} \mid x_1, \dots, x_k)$ can be determined analytically.
In Section 3.4.1 we formalize the concept of conjugacy, introduce the exponential family distribution and give two basic examples.
In Section 3.4.2 we show some basic properties of exponential family distributions, which we need in Section 3.4.4.
The existence and formula of a conjugate prior distribution in case of exponential family distributions is shown in Section 3.4.3. Furthermore, it is shown that also the predictive distribution can be determined analytically.
In case of an exponential family distribution and its standard conjugate prior, cf. Section 3.4.3, we have posterior linearity in the expectation of the sufficient statistic, see Section 3.4.4. Furthermore, under some regulatory conditions also the opposite holds for exponential family distributions: If we have posterior linearity in the sufficient statistic, then the prior distribution must be the standard conjugate prior (up to a reparameterization).

3.4.1 Exponential family distribution

Definition 3.4.1. *Consider a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$. A family $\mathcal{G}$ of prior distributions $\Pi(\cdot)$ is called conjugate for a sampling model $\mathcal{P} = \{P_\theta : \theta \in \Omega_\theta\}$, if for any prior distribution $\Pi \in \mathcal{G}$ and any $\theta \in \Omega_\theta$ the corresponding posterior distribution $\Pi(\cdot \mid X = x)$, for $X \sim P_\theta$, is in the same class $\mathcal{G}$, i.e.,*

$$\Pi(\cdot) \in \mathcal{G} \Rightarrow \Pi(\theta \mid X = x) \in \mathcal{G}.$$

Definition 3.4.2. *A family $\mathcal{E}$ of probability distributions is a d-dimensional exponential family, if the density (or probability mass function) of any member of $\mathcal{E}$ has the general form*

$$p_\theta(x) = h(x) \cdot \exp\left(\theta^T t(x) - A(\theta)\right),$$

for a parameter vector $\theta = (\theta_1, \dots, \theta_d)^T \in \mathbb{R}^d$, the canonical parameter, a non-negative function $h : \mathbb{R}^k \to \mathbb{R}$ and a function $t : \mathbb{R}^k \to \mathbb{R}^d$. Furthermore, the function $A(\theta)$, denoted as the cumulant function, is given by

$$A(\theta) = \log \int h(x) \exp\left(\theta^T t(x)\right) dx$$

*and can be viewed as the logarithm of a normalization factor.
The set of parameters θ for which the integral is finite is referred to as the natural parameter space*

$$\mathcal{N} := \left\{\theta : \int h(x) \exp\left(\theta^T t(x)\right) dx < \infty\right\}.$$

If the natural parameter space is a non-empty open set, the family $\mathcal{E}$ is called regular.

Remark 3.4.3. *According to the Fisher-Neyman factorization theorem, which states that if the likelihood function is $p_\theta(\cdot)$, then $t(\cdot)$ is sufficient for θ if and only if non-negative functions $g(\cdot)$ and $h(\cdot)$ can be found such that*

$$p_\theta(x) = h(x) g_\theta(t(x)),$$

we obtain for exponential family distributions, that $t(x) := (t_1(x), \dots, t_d(x))^T$ is a sufficient statistic for θ.

Remark 3.4.4. *$A(\theta)$ is not a degree of freedom in the specification of an exponential family density, it is determined once $t(x)$ and $h(x)$ are determined.*

Remark 3.4.5. *We will restrict only on regular exponential families.*

Example 3.4.6. *Let $X \sim Ber(\eta)$. Then we rewrite the probability weights by*

$$\begin{aligned} p_\eta(x) &= \eta^x (1-\eta)^{1-x} \\ &= \exp(x \log \eta + (1-x)\log(1-\eta)) \\ &= \exp\left(\log\left(\frac{\eta}{1-\eta}\right) x + \log(1-\eta)\right), \end{aligned}$$

thus the Bernoulli distribution is an exponential family distribution with

$$\theta = \log \frac{\eta}{1-\eta}; \quad h(x) = 1; \quad t(x) = x; \quad A(\theta) = -\log(1-\eta) = \log\left(1+e^\theta\right).$$

Example 3.4.7. *Let $N \sim Pois(\lambda)$. Rewriting the probability mass function we obtain*

$$\begin{aligned} p_\lambda(x) &= \frac{\lambda^n}{n!}\exp(-\lambda) \\ &= \frac{1}{n!}\exp(n\log\lambda - \lambda), \end{aligned}$$

thus the Poisson distribution is an exponential family distribution with

$$\theta = \log \lambda; \quad h(n) = \frac{1}{n!}; \quad t(n) = n; \quad A(\theta) = \lambda = e^\theta.$$

3.4.2 Mean and variance of the sufficient statistic

In Definition 3.4.2 we have already denoted $A(\theta)$ as the cumulant function. We will first motivate this identification with an example of the Bernoulli distribution and afterwards we will show that for exponential family distributions cumulants can be determined by calculating derivatives of $A(\theta)$.

Example 3.4.8. *Let $X \sim Ber(\eta)$ as in Example 3.4.6. We have already seen, that $A(\theta) = \log\left(1+e^\theta\right)$ with $\theta = \log \frac{\eta}{1-\eta}$. Taking a first derivative yields*

$$\frac{\partial A(\theta)}{\partial \theta} = \frac{e^\theta}{1+e^\theta} = \frac{1}{1+e^{-\theta}} = \frac{1}{\frac{1-\eta}{\eta}} = \eta,$$

which is the mean of the Bernoulli distribution or in other words the mean of the sufficient statistic $t(x) = x$.
Taking a second derivative yields

$$\begin{aligned} \frac{\partial^2 A(\theta)}{\partial \theta^2} &= \frac{\partial}{\partial \theta}\frac{1}{1+e^{-\theta}} = \frac{1}{(1+e^{-\theta})^2}e^{-\theta} \\ &= \eta^2 \frac{1-\eta}{\eta} = \eta(1-\eta), \end{aligned}$$

which is the variance of the Bernoulli distribution or again the variance of the sufficient statistic.

To verify the results of example 3.4.8 in the general setting of an exponential family distributions, we need the following technical Lemma:

Lemma 3.4.9. *As in Definition 3.4.2 let $h(\cdot)$ be a non-negative function and let $\mathcal{N}_f$ be the set of values of $\theta \in \mathbb{R}^d$ where*

$$\int |f(x)| h(x) \exp\left(\theta^T t(x)\right) dx < \infty.$$

Then the function

$$g(\theta) = \int f(x) h(x) \exp\left(\theta^T t(x)\right) dx$$

is continuous and has continuous partial derivatives of all orders of θ in the interior of $\mathcal{N}_f$. Furthermore, these derivatives can be computed by differentiation under the integral sign.

Proof. See Brown (1986). Basically the proof relies on the dominated convergence theorem. □

Lemma 3.4.10. *Let $\mathcal{E}$ be a regular exponential family distribution and*

$$A(\theta) = \log \int h(x) \exp\left(\theta^T t(x)\right) dx.$$

Then the mean of the sufficient statistic can be obtained by computing a first derivative of the cumulant function $A(\theta)$ and the variance by the second derivative of $A(\theta)$, i.e.,

$$\frac{\partial A(\theta)}{\partial \theta} = \mathbb{E}[t(X)] \qquad \text{and} \qquad \frac{\partial^2 A(\theta)}{\partial \theta^2} = \mathbb{V}ar[t(X)].$$

Proof. According to Lemma 3.4.9 with $f = 1$ and the assumption of a regular family, we can calculate the derivatives of $A(\theta)$ by differentiation under the integral sign. Thus, we obtain for the first derivative of $A(\theta)$

$$\begin{aligned}
\frac{\partial A(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \log \int h(x) \exp\left(\theta^T t(x)\right) dx \\
&= \frac{\int \frac{\partial}{\partial \theta} h(x) \exp\left(\theta^T t(x)\right) dx}{\int h(x) \exp\left(\theta^T t(x)\right) dx} \\
&= \int t(x) h(x) \exp\left(\theta^T t(x)\right) dx \exp(-A(\theta)) \\
&= \int t(x) h(x) \exp\left(\theta^T t(x) - A(\theta)\right) dx \\
&= \mathbb{E}[t(X)].
\end{aligned}$$

Similarly calculating the second derivative yields

$$\begin{aligned}
\frac{\partial^2 A(\theta)}{\partial \theta^2} &= \int t(x) \left(t(x) - \frac{\partial A(\theta)}{\partial \theta} \right)^T h(x) \exp\left(\theta^T t(x) - A(\theta)\right) dx \\
&= \int t(x) \left(t(x) - \mathbb{E}[t(X)]\right)^T h(x) \exp\left(\theta^T t(x) - A(\theta)\right) dx \\
&= \mathbb{E}[t(X) t(X)^T] - \mathbb{E}[t(X)]\mathbb{E}[t(X)]^T \\
&= \mathbb{V}ar[t(X)].
\end{aligned}$$

□

Remark 3.4.11. *If the exponential family distribution is presented in its so called mean parametrization, i.e., $t(x) = x$, cf. Example 3.4.8, then the first derivative of $A(\theta)$ yields the expectation of the distribution and the second derivative yields the variance.*

Remark 3.4.12. *For samples of sizes $k \in \mathbb{N}$ the sufficient statistic*

$$t_n(x) = \sum_{i=1}^{k} t(x_i)$$

is a sum of independent random variables, so by the Central Limit Theorem we have approximately

$$t_n(X) \sim \mathcal{N}\left(k\frac{\partial A(\theta)}{\partial \theta}, k\frac{\partial^2 A(\theta)}{\partial \theta^2}\right).$$

3.4.3 Existence of conjugate prior distribution and posterior updating

In case of an exponential family distribution there exists a conjugate prior distribution and its structure can be easy obtained by mimicking the likelihood of the sample distribution, cf. Lemma 3.4.13. By the example of the Bernoulli distribution, we use Lemma 3.4.13 to determine a conjugate prior distribution and illustrate the prior parameters, which in general can interpreted as prior sample size and prior belief. If we choose appropriate prior parameters, see Lemma 3.4.16, the prior distribution is normalizable and also the predictive distribution can be determined analytically, see Lemma 3.4.17.

Lemma 3.4.13. *If the sampling model is a regular d-dimensional exponential family with density (or probability mass function)*

$$p_\theta(x) = h(x) \cdot \exp\left(\theta^T t(x) - A(\theta)\right),$$

then there exists a conjugate prior distribution $\pi_\theta(\cdot)$ and can be obtained by mimicking the likelihood

$$\pi(\theta) \sim \exp(\theta^T (k_0 t_0) - k_0 A(\theta)), \tag{3.6}$$

where $k_0 > 0$ and $t_0 \in \mathbb{R}^d$. This prior distribution is called the standard conjugate prior. The posterior distribution is of the form

$$\pi(\theta \mid X_{1:k} = x_{1:k}) \sim \exp(\theta^T (k_0 t_0 + \sum_{i=1}^{k} t(x_i)) - (k_0 + k)A(\theta)).$$

Proof. Consider the likelihood of an i.i.d. sample $x_1, \dots, x_k$

$$\begin{aligned} p_\theta(x_1, \dots, x_k) &= \prod_{i=1}^{k} p_\theta(x_i) = \prod_{i=1}^{k} h(x_i) \exp\left(\theta^T t(x_i) - A(\theta)\right) \\ &= \left(\prod_{i=1}^{k} h(x_i)\right) \exp\left(\theta^T \left(\sum_{i=1}^{k} t(x_i)\right) - kA(\theta)\right). \end{aligned}$$

To see, that Equation (3.6) is a conjugate prior, we determine the posterior density

$$\begin{aligned}\pi(\theta \mid X_{1:k} = x_{1:k}) &\sim \pi(\theta) \cdot p_\theta(x_1, \ldots, x_k) \\ &\sim \exp(\theta^T(k_0 t_0) - k_0 A(\theta)) \cdot \exp\left(\theta^T \left(\sum_{i=1}^k t(x_i)\right) - kA(\theta)\right) \\ &\sim \exp(\theta^T(k_0 t_0 + \sum_{i=1}^k t(x_i)) - (k_0 + k)A(\theta)),\end{aligned}$$

which retains the form of Equation (3.6) and thus is in the same family as the prior. □

Remark 3.4.14. *The prior to posterior conversion can be summarized with the following update rules:*

$$\begin{aligned}k_0 &\to k_0 + k \\ k_0 t_0 &\to k_0 t_0 + \sum_{i=1}^k t(x_i),\end{aligned}$$

where k_0 can be interpreted as prior sample size and t_0 as prior belief. If we have a strong opinion t_0 about the real value of the unknown θ, we choose k_0 large. If we are wrong, a large sample size k is needed to correct our wrong prior belief by the observed information $\sum_{i=1}^k t(x_i)$.

Example 3.4.15. *Using the results of Example 3.4.6, we obtain from Lemma 3.4.13 for the conjugate prior distribution*

$$\pi(\theta) \sim \exp\left(\theta(k_0 t_0) - k_0 \log\left(1 + e^\theta\right)\right).$$

Transforming back to the usual parametrization η we write

$$\theta = \log \frac{\eta}{1-\eta} =: g^{-1}(\eta).$$

To transform the prior distribution we use the change of variable formula

$$\pi(\theta) = \left|\frac{\partial g^{-1}(\eta)}{\partial \eta}\right| \pi(g^{-1}(\eta)).$$

Calculating the first derivative of $g^{-1}(\cdot)$ yields

$$\frac{\partial g^{-1}(\eta)}{\partial \eta} = \frac{1-\eta}{\eta} \frac{(1-\eta)+\eta}{(1-\eta)^2} = \frac{1}{\eta(1-\eta)}.$$

Thus, we get for the variable transformation

$$\begin{aligned}\pi(\theta) &\sim \frac{1}{\eta(1-\eta)} \exp\left(\log\left(\frac{\eta}{1-\eta}\right)(k_0 t_0) - k_0 \log\left(1 + \frac{\eta}{1-\eta}\right)\right) \\ &= \frac{1}{\eta(1-\eta)} \left(\frac{\eta}{1-\eta}\right)^{k_0 t_0} \left(1 + \frac{\eta}{1-\eta}\right)^{-k_0} \\ &= \frac{1}{\eta(1-\eta)} \left(\frac{\eta}{1-\eta}\right)^{k_0 t_0} (1-\eta)^{k_0} \\ &= \eta^{k_0 t_0 - 1}(1-\eta)^{k_0(1-t_0)-1} \sim Beta(k_0 t_0, k_0(1-t_0)),\end{aligned}$$

what we have already seen in Section 3.2.1. In Figure 3.3 we can nicely observe the interpretation of the prior parameters k_0 and t_0: Thereby k_0 can be interpreted as prior sample size and represents our confidence in the prior belief t_0, which equals the prior belief about η, as $t(x) = x$ is a sufficient statistic for η.

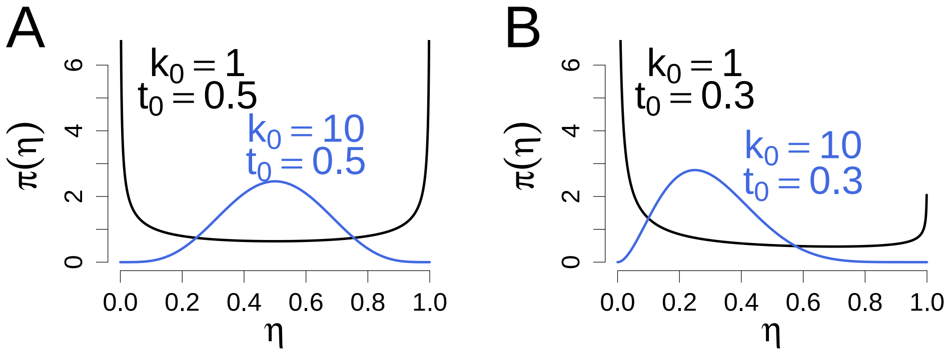


Figure 3.3: Interpretation of the prior parameters k_0 and t_0 by the example of a Bernoulli sample distribution, cf. Example 3.4.15. A. We choose a prior belief of $t_0 = 0.5$. B. We choose a prior belief of $t_0 = 0.3$. The black line represents a small prior sample size of $k_0 = 1$, the blue line a prior sample size of $k_0 = 10$, where the density peaks clearly at the prior belief t_0.

Even if Lemma 3.4.13 shows us the structure of a conjugate prior, this can not always help us in practice. We need to be able to calculate the normalization factor, which is not possible in general. However in case of the usual distributions, it helps to determine appropriate prior distributions with a definite procedure, see Section 4.5. The following Lemma states in which cases Equation (3.6) yields a normalizable distribution:

Lemma 3.4.16. *The conjugate prior distribution 3.6 is normalizable if and only if $k_0 > 0$ and t_0 lies in the interior of the convex hull of the support of the measure $\mu(dx) = h(x)dx$.*

Proof. See Diaconis and Ylvisaker (1985). □

Another useful property of an exponential family distribution marginalized over its standard conjugate prior distribution is, that if the prior distribution is normalizable and we know the explicit expression, the probability density function of the predictive distribution can be determined analytically, see the following lemma:

Lemma 3.4.17. *Let us consider a regular d-dimensional exponential family model with density*

$$p_\theta(x) = h(x) \cdot \exp\left(\theta^T t(x) - A(\theta)\right),$$

and a conjugate prior distribution which is normalizable, i.e.,

$$\pi(\theta) = g(k_0, t_0) \exp(\theta^T (k_0 t_0) - k_0 A(\theta)).$$

Then the predictive distribution can be determined analytically by

$$p_\theta(x_{k+1} \mid X_{1:k} = x_{1:k}) = h(x_{k+1}) \frac{g\left(\tilde{k}, \tilde{t}\right)}{g\left(\tilde{k} + 1, \frac{\tilde{k}\tilde{t} + t(x_{k+1})}{\tilde{k}+1}\right)},$$

where

$$\tilde{k} := k_0 + k \quad \textit{and} \quad \tilde{t} := \frac{k_0 t_0 + \sum_{i=1}^{k} t(x_i)}{k_0 + k}.$$

Proof. Out of Lemma 3.4.13 we know how to get the posterior distribution by simply updating the parameters of the prior distribution. Thus also for the predictive distribution it holds

$$p_\theta(x_{k+1} \,|\, X_{1:k} = x_{1:k}) = p_\theta\left(x_{k+1} \,|\, \tilde{k}, \tilde{t}\right),$$

so if we know the structure of the distribution, we only need to update the parameters k_0 and t_0. Hence

$$\begin{aligned} p_\theta(x_{k+1} \,|\, X_{1:k} = x_{1:k}) &= \int p_\theta(x)\pi(\theta \,|\, \tilde{k}, \tilde{t}) d\theta \\ &= \int h(x_{k+1}) \cdot \exp\left(\theta^T t(x_{k+1}) - A(\theta)\right) g(\tilde{k}, \tilde{t}) \exp\left(\theta^T(\tilde{k}\tilde{t}) - \tilde{k}A(\theta)\right) \\ &= h(x_{k+1}) g(\tilde{k}, \tilde{t}) \int \exp\left(\theta^T(t(x_{k+1}) + \tilde{k}\tilde{t}) - (\tilde{k}+1)A(\theta)\right) \\ &= h(x_{k+1}) \frac{g\left(\tilde{k}, \tilde{t}\right)}{g\left(\tilde{k}+1, \frac{\tilde{k}\tilde{t}+t(x_{k+1})}{\tilde{k}+1}\right)} \end{aligned}$$

□

3.4.4 Posterior linearity

In Section 3.2.3 we have seen that in case of a Bernoulli distribution the posterior expectation of Θ is a convex combination of the prior parameter θ_0 and the maximum likelihood estimator $\hat{\theta}_{ML}$. Here we will see, cf. Theorem 3.4.20, that in case of an exponential family distribution and its standard conjugate prior distribution, cf. Lemma 3.4.13, posterior linearity holds for the sufficient statistic $t(X)$ in general. As in case of a Bernoulli distribution $t(X) = X$, cf. Example 3.4.6, and

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \,|\, \Theta]] = \mathbb{E}[\Theta]$$

and the Beta prior distribution can be obtained by variable transformation from the standard prior, see Example 3.4.15, its just a special case of Theorem 3.4.20.
Under some regulatory conditions also the opposite holds for exponential family distributions: If we have posterior linearity in the sufficient statistic, then the prior distribution of the canonical parameter must be the standard conjugate prior.

Lemma 3.4.18. *Consider a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ with P_θ an exponential family distribution and let Π its normalizable standard prior distribution, i.e.,*

$$\pi(\theta) = g(k_0, t_0) \exp\left(\theta^T(k_0 t_0) - k_0 A(\theta)\right).$$

Then it holds for $\Theta \sim \Pi$

$$\mathbb{E}\left[\frac{\partial A(\Theta)}{\partial \Theta}\right] = t_0.$$

Proof. First we observe that

$$\frac{\partial \pi(\theta)}{d\theta} = k_0 \left(t_0 - \frac{\partial A(\theta)}{d\theta} \right) \pi(\theta).$$

According to Lemma 3.4.9

$$\int \frac{\partial \pi(\theta)}{\partial \theta} \, d\theta = \frac{\partial}{d\theta} \int \pi(\theta) \, d\theta = 0$$

and as

$$t_0 \int \pi(\theta) \, d\theta = t_0,$$

we obtain overall

$$\begin{aligned} \mathbb{E}\left[\frac{\partial A(\Theta)}{\partial \Theta} \right] &= \int \frac{\partial A(\theta)}{\partial \theta} \pi(\theta) \, d\theta \\ &= t_0 - \int \left(t_0 - \frac{\partial A(\theta)}{\partial \theta} \right) \pi(\theta) \, d\theta \\ &= t_0 - \frac{1}{k_0} \int \frac{\partial \pi(\theta)}{\partial \theta} \, d\theta \\ &= t_0. \end{aligned}$$

□

Lemma 3.4.19. *Consider a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ with P_θ an exponential family distribution, i.e.,*

$$p_\theta(x) = h(x) \cdot \exp\left(\theta^T t(x) - A(\theta) \right),$$

and let Π its normalizable standard prior distribution. Let X be generated by $\mathcal{B}(\Pi, \mathcal{P})$, then it holds

$$\mathbb{E}\left[t(X) \right] = t_0.$$

Proof. Combining Lemma 3.4.10 and Lemma 3.4.18 yields

$$\mathbb{E}\left[t(X) \right] = \mathbb{E}\left[\mathbb{E}\left[t(X) \mid \Theta \right] \right] = \mathbb{E}\left[\frac{\partial A(\Theta)}{\partial \Theta} \right] = t_0$$

□

Theorem 3.4.20. *Consider a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ with P_θ an exponential family distribution and let Π its normalizable standard prior distribution. Let $X_1, \ldots, X_k$ be generated by $\mathcal{B}(\Pi, \mathcal{P})$, then for the posterior expectation of the sufficient statistic holds*

$$\mathbb{E}\left[t(X) \mid X_{1:k} = x_{1:k} \right] = \frac{k_0 t_0 + k_0 \bar{t}(x)}{k_0 + k},$$

with $\bar{t}(x) := 1/k \sum_{i=1}^{k} t(x_i)$.

Proof. According to Lemma 3.4.13 the posterior distribution of Θ is

$$\pi(\theta \mid X_{1:k} = x_{1:k}) \sim \exp(\theta^T(k_0 t_0 + \sum_{i=1}^{k} t(x_i)) - (k_0 + k)A(\theta)).$$

Applying Lemma 3.4.19 on the posterior distribution gives the statement. □

Remark 3.4.21. *The posterior expectation of the sufficient statistic of θ can be written as*

$$\mathbb{E}\left[t(X) \mid X_{1:k} = x_{1:k}\right] = ct_0 + (1-c)\bar{t}(x), \quad c := \frac{k_0}{k_0 + k},$$

which is a convex combination of the prior expectation t_0 and mean observation $\bar{t}(x)$. The weights are proportional to k_0 and the sample size k, so we see again the interpretation of k_0 as prior sample size.

Remark 3.4.22. *The posterior expectation can be also described as a convex combination of the prior expectation t_0 and the maximum likelihood estimator $\hat{\theta}_{ML}$, since*

$$\ell(\theta \mid x_1, \ldots, x_k) := \log(p_\theta(X_{1:k} = x_{1:k})) \sim \theta^T \sum t(x_i) - kA(\theta).$$

Taking the derivative with respect to θ yields

$$\frac{\partial \ell(\theta \mid X_{1:k} = x_{1:k})}{\partial \theta} = \sum_{i=1}^{k} t(x_i) - k\frac{\partial A(\theta)}{\partial \theta}$$

and setting to zero gives

$$\begin{aligned} \frac{\partial A(\theta)}{\partial \theta} &= \frac{1}{k}\sum_{i=1}^{k} t(x_i) \\ \Longrightarrow \qquad \hat{\theta}_{ML} &= \frac{\partial A(\theta)^{-1}}{\partial \theta}\left(\sum_{i=1}^{k} t(x_i)\right), \end{aligned}$$

which exists as, recall Lemma 3.4.10, $\mathbb{E}[t(X)] = \frac{\partial A(\theta)}{\partial \theta}$ is a strictly monotone increasing function $\left(\frac{\partial^2 A(\theta)}{\partial \theta} = \mathbb{V}ar[t(X)] > 0\right)$.

In case of an exponential family distribution given in its canonical parametrization and if we choose the standard conjugate prior distribution, we know according to Theorem 3.4.20, that posterior linearity in the expectation of the sufficient statistic always holds. But often the standard exponential families are parameterized in other terms, i.e., the parametrization involving the success probability η for the binomial distribution.

In the introductory example we have seen that posterior linearity even holds for this parametrization. The reason can be recognized in Example 3.4.15: The transformation of the standard conjugate prior in its canonical parametrization to the usual parametrization can be obtained by using the change-of-variable formula, i.e for $\theta = g^{-1}(\eta)$

$$\begin{aligned} \pi(\eta) &= \left|\frac{\partial g^{-1}(\eta)}{\partial \eta}\right| \pi\left(g^{-1}(\eta)\right) \\ &\sim \left|\frac{\partial g^{-1}(\eta)}{\partial \eta}\right| \exp\left(g^{-1}(\eta)^T(k_0 t_0) - k_0 A\left(g^{-1}(\eta)\right)\right). \end{aligned}$$

As we simply change variables, we still get posterior linearity. Descriptive spoken the Jacobian factor simply ensures the additional property of posterior linearity, as the following remark addresses.

Remark 3.4.23. *Another family of conjugate prior is given by*

$$\pi(\eta) = \exp\left(g^{-1}(\eta)^T(k_0t_0) - k_0A\left(g^{-1}(\eta)\right)\right),$$

without the Jacobian factor. In general these two families of conjugate priors are not identical. But in case of natural exponential families ($t(x) = x$) the two families are identical (up to a reparameterization) if and only if the exponential family is quadratic (the variance of the distribution is a quadratic polynomial in the mean), i.e., for $\mu := \mathbb{E}[X]$

$$\mathbb{V}ar[X] = \nu_0 + \nu_1\mu + \nu_2\mu^2, \quad \nu_0, \nu_1, \nu_2 \in \mathbb{R}.$$

Examples for natural exponential families with quadratic variance function are the Normal distribution, Poisson, Gamma, Binomial and Negative Binomial distribution. Further information can be found in Consonni and Veronese (1992); Gutiérrez-Pena and Smith (2003, 1995).

Theorem 3.4.24. *Consider a Bayesian model $\mathcal{B}(\Pi, \mathcal{P})$ with P_θ a continuous exponential family distribution. Let X_1, X_2 be generated by $\mathcal{B}(\Pi, \mathcal{P})$ and suppose the support of the measure $\mu(dx) = h(x)dx$ contains an open interval in $\mathbb{R}^d$. If Θ has a prior distribution Π, which is not concentrated at a single point and if*

$$\mathbb{E}[t(X_2) \mid X_1] = a \cdot t(X_1) + b,$$

for some constant $a \in \mathbb{R}$ and $b \in \mathbb{R}^d$, then $a \neq 0$ and the prior density $\pi(\cdot)$ is given by

$$\pi(\theta) \sim \exp\left(a^{-1}\theta^T b - a^{-1}(1-a)A(\theta)\right).$$

Proof. See Diaconis and Ylvisaker (1979). □

Remark 3.4.25. *The result above even holds for any usual discrete family and a version of Theorem 3.4.24 appropriate for discrete data is also given in Diaconis and Ylvisaker (1979).*

Remark 3.4.26. *We can easily transfer the form of the prior distribution in Theorem 3.4.24 to the standard conjugate prior: We can find k_0, t_0 such that*

$$\mathbb{E}[t(X_2) \mid X_1] = \underbrace{\frac{1}{k_0+1}}_{a} \cdot t(X_1) + \underbrace{\frac{k_0}{k_0+1}}_{b} t_0,$$

which yields

$$a^{-1}b = \left(\frac{1}{k_0+1}\right)^{-1} \frac{k_0}{k_0+1} t_0 = k_0t_0$$

and

$$a^{-1}(1-a) = (k_0+1)(1 - 1/(k_0+1)) = k_0.$$

Thus we obtain the standard exponential conjugate prior

$$\pi(\theta) \sim \exp\left(\theta^T(k_0t_0) - k_0A(\theta)\right).$$

Chapter 4

Detection of changes in the stimulus

In this section we investigate whether imprecise phases can improve the detection of *changes* in the stimulus. Results are obtained on basis of one neuron, for multiple neurons see Section 5.3.3. To this end, we extend a simplified version of our GLO-model (Section 1.1) for $M = 1$ neuron by assuming a sequence of oscillation cycles with deterministic and known length and assuming that the assignment of spikes to a particular oscillation cycle is known, where the rate and phase parameters can change between cycles (Section 4.1). We particularly investigate the number of correctly (the distance to a true change point is at most three) and of falsely detected change points in the bivariate analysis in which rate and phase parameters are assumed to change simultaneously, as compared to the approach in which changes in rate or phase parameters are analyzed individually.

Change point detection is performed using a Bayesian online change point detection algorithm (BOCD) (Adams and MacKay, 2007) (Section 4.2), which includes the estimation of the change point rate (Wilson et al., 2010) (Section 4.2.2). This algorithm directly matches the assumptions of the change point model. Given the model assumptions, the BOCD derives exactly the probability of a change point for a given oscillation cycle. The algorithm includes prior information on neurophysiologically plausible parameter ranges as well as information about distributional assumptions. As a consequence, it can operate on small time scales such as a few oscillation cycles, in contrast to asymptotic methods. In addition, the BOCD is capable of estimating the rate with which change points are observed, and computational speed is increased by constant online updating of the derived probabilities.

The BOCD assumes a prior distribution for rate and phase parameters, for which we choose conjugate prior distributions for convenience (phase in Section 4.3.1 and rate in Section 4.3.2). As the BOCD crucially depends on the choice of the prior parameters, we discuss parameter choice for the relevant parameter range, see Sections 4.3.1.2 (phase and change point prior) and 4.3.2.2 (rate) and 4.3.3.2 (rate and phase choice in bivariate analysis).

Section 4.3.3 then investigates the number of correctly and falsely detected change points for a pure rate or pure phase code and the improvement in the bivariate analysis when rate and phase parameters are assumed to change simultaneously. While the BOCD evaluates the occurrence of change points at the end of the time series, we investigate in Section 4.2.4 an extension which works literally 'on-line', in which decisions about a change need to be made ad hoc or after a very small number of oscillation cycles, and investigate its improvement in the bivariate case.

Afterwards in Section 4.4 we discuss an approach which allows to consider changes in rate and phase as dependent but still provides an efficient calculation by conjugate distributions. Thus we are able to consider special information about stimuli properties: For example we have special knowledge about the stimuli structure and know that stimuli are coded either by small rates or by high rates, but not by middle rates. Or we know that in a stimulus only small rates are connected with small phases or high rates with high phases.

In the thesis we mostly assume σ (the precision of the spike timing) is fixed and known. In Section 4.5 we extend this assumption by an unknown and random spike time precision ς, which is locked to the global change point process, i.e., we assume changes in rate, phase and precision occur simultaneously according to the prior distribution. Our aim is to explore, how a change in the spiking precision impacts the change point detection in case of a pure phase analysis. Therefore we first investigate how changes in the spiking precision affect the ability of the phase to detect change points, if we erroneously assume a constant spiking precision σ^2. Second, we consider changes in the precision, phase and rate occur simultaneously and consider the benefit of a simultaneous analysis of a change in the precision, phase and rate compared to a pure rate analysis.

4.1 Change point model

We assume the following change point model: First we assume that change points occur independently and with equal probability η for every oscillation cycle. Formally, let $Y_1, Y_2, \dots$ denote a sequence of independent Bernoulli random variables with $\mathbb{P}(Y_1 = 1) = \eta$, where $Y_k = 1$ indicates that a change point occurs at cycle k. Second, let $\Lambda_0, \Lambda_1, \dots$ be a sequence of i.i.d. rate parameters with prior distribution $\pi_\lambda(\cdot)$ and $\lambda_0, \lambda_1, \dots$ a random realization. Similarly, let $\Phi_0, \Phi_1, \dots$ be a sequence of i.i.d. phase parameters with prior distribution $\pi_\varphi(\cdot)$ and $\varphi_0, \varphi_1, \dots$ a random realization. At every change point, a new realization of Λ, Φ is drawn. As a consequence, let $A_k := \sum_{i=1}^{k} Y_i$ denote the number of change points up to time k, where we set $A_0 := 0$. Then in cycle k, overall $N_k \sim Pois(\lambda_{A_k})$ spikes are chosen and are placed independently according to a $\mathcal{N}(\varphi_{A_k}, \sigma^2)$-distribution (the spike times are denoted by $X_1^{(k)}, \dots, X_{N_k}^{(k)}$), where we assume the precision of the spike timing σ to be fixed and known. Again we set $\sigma = 1$ and scale the phase range accordingly.

A graphical description of the change point model is shown in Figure 4.1. We use the same spike generating process as in Section 2.1, but now we observe a sequence of oscillation cycles. Again, we artificially assume that the reference time of an oscillation cycle is known as well as the assignment of each spike to its respective oscillation cycle.

Note that we first assume that rate and phase parameters change independently at a change point, so it would be appropriate to write $\pi_\lambda \times \pi_\varphi$ in Figure 4.1, but later we will discuss a procedure to consider dependencies between rate and phase, see Section 4.4.

4.2 Bayesian Online Change Point Detection Algorithm

Here we summarize the BOCD proposed in (Adams and MacKay, 2007), including an extension (Wilson et al., 2010) (Section 4.2.2) in which the change point probability η of the change point process is estimated. We first use a general notation with general parameters θ, before replacing this notation by the specific rate and phase parameters in Section 4.3.

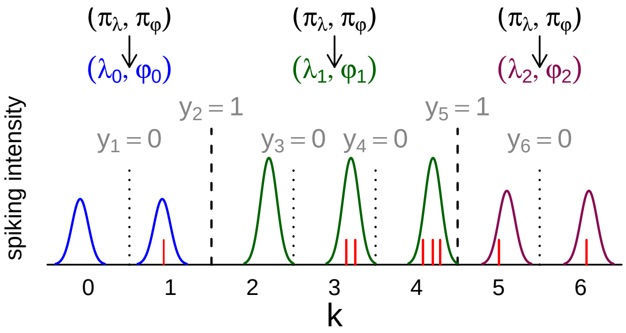


Figure 4.1: Change point model: The activity in every oscillation cycle is assumed to follow an inhomogeneous Poisson process as in Section 2.1, i.e., the number of spikes per cycle is assumed Poisson distributed with the respective rate parameter, and the spike time is assumed independent and normally distributed with variance 1 and mean given by the respective phase parameter. Rate and phase parameters can change in time as follows. For each oscillation cycle an independent Bernoulli random variable $Y_k \sim Ber(\eta)$ indicates whether a change point occurs. If no change point occurs ($Y_k = 0$), the rate λ and phase φ remain identical to the previous oscillation cycle ($k-1$). If a change point occurs ($Y_k = 1$), new parameters for rate λ and phase φ are chosen independently according to the prior distributions π_λ and π_φ.

Let $\Theta_0, \Theta_1, \ldots$ be a sequence of i.i.d parameters and let $\theta_0, \theta_1, \ldots$ be a random realization. In cycle k we observe a random variable X_k, whose distribution $p_\theta(\cdot)$ depends on the parameter θ . The distribution of $X_{0:k} := (X_0, \ldots, X_k)$ can be expressed as

$$p(x_{0:k} \,|\, (A_1 = a_1, \ldots, A_k = a_k), (\Theta_0 = \theta_0, \ldots, \Theta_k = \theta_k)) = p_{\theta_0}(x_0) \prod_{i=1}^{k} p_{\theta_{a_i}}(x_i).$$

This representation makes use of the fact that the sequence of observations up to time k can be divided into $A_k + 1$ segments, where each segment i contains i.i.d. observations with parameter θ_{a_i}.

4.2.1 The BOCD: Change points encoded in run length

The BOCD is based on the *run length* r_k, which represents the time since the last change point and is a direct function of the set of change points and the cycle k (Figure 4.2 A). If no change point occurs at cycle k, the run length increases by 1 at cycle k, otherwise it drops to 0. Formally, the run length R_k at time k is a random variable given by

$$R_k := \begin{cases} \min(i \geq 0 : \; A_k - A_{k-1-i} = 1), & \text{if } A_k > 0, \\ k, & \text{else.} \end{cases}$$

If we know all run lengths up to time K, we know the positions of all change points. The objective of the BOCD is therefore to estimate the run length at every cycle k. Therefore we need to consider all possible run length paths, cf. Figure 4.2 B. This is done recursively (see Figure 4.2 C gray dots). At the last cycle K, the most likely run length is chosen and used to estimate the change points backwards in time. In Figure 4.2 C this results in two detected

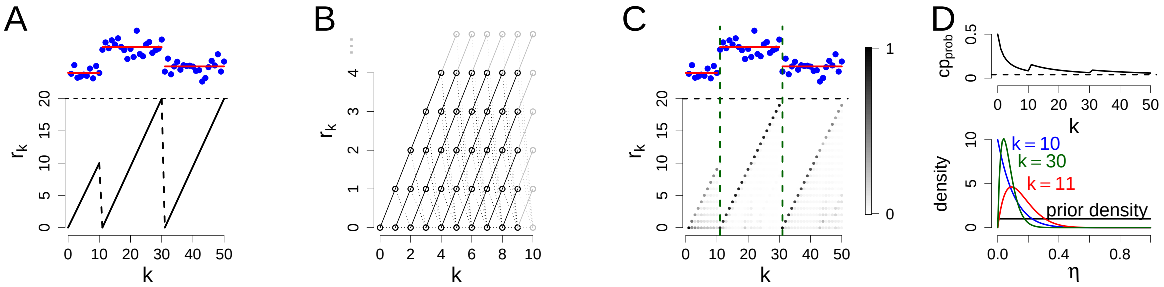


Figure 4.2: A. An example sequence of random variables with two change points in the mean and the corresponding theoretical run lengths r_k. B. All possible run length paths up to time $K = 10$. C. The same sequence of random variables as in A and the estimated distributions of run lengths, cf. Equation (4.1), for every cycle k, illustrated with a color code. The estimated change points are shown by the vertical dashed green lines, which result of the path with maximal posterior probability at time $K = 50$. D. Illustration of the predictive change point probability $cp_{prob} = \mathbb{P}(Y_k = 1 \,|\, A_{k-1} = a_{k-1})$ for the change point path estimated in C (Equation (4.2)) (solid), converging to the real change point probability (dashed).

change points (green dashed lines). Note that because the algorithm first evaluates the run lengths at time K, it cannot be interpreted rigorously as an 'online' procedure. A modified algorithm that uses only information up to the current cycle or only a few cycles in the future is investigated in Section 4.2.4.

In detail, in order to estimate the run lengths, one needs to derive

$$\mathbb{P}(R_k = r_k \,|\, X_{0:k} = x_{0:k}) = \frac{p(x_{0:k} \,|\, R_k = r_k)\mathbb{P}(R_k = r_k)}{p(x_{0:k})} \qquad \forall\, r_k \in 0, \ldots, k, \tag{4.1}$$

where the denominator is a normalization factor that can be neglected. The numerator can be calculated recursively. For better reading we introduce the notation $x^{(r_k)} := (x_{k-1}, \ldots, x_{k-r_k})$ indicating the set of observations associated with the run length r_k (note that this set can be empty):

$$\begin{aligned}
&\mathbb{P}(R_k = r_k)\, p(x_{0:k} \,|\, R_k = r_k) \\
&= \sum_{i=0}^{k-1} \mathbb{P}(R_k = r_k, R_{k-1} = i)\, p(x_{0:k} \,|\, R_k = r_k, R_{k-1} = i) \\
&= \sum_{i=0}^{k-1} \mathbb{P}(R_k = r_k \,|\, R_{k-1} = i)\, \mathbb{P}(R_{k-1} = i)\, p(x_k \,|\, R_k = r_k, R_{k-1} = i, x_{0:k-1})\, p(x_{0:k-1} \,|\, R_{k-1} = i) \\
&= \sum_{i=0}^{k-1} \mathbb{P}(R_k = r_k \,|\, R_{k-1} = i)\;\, p(x_k \,|\, x^{(r_k)})\, p(x_{0:k-1} \,|\, R_{k-1} = i)\mathbb{P}(R_{k-1} = i).
\end{aligned}$$

This sum reduces to one summand $i = r_k - 1$ if $r_k > 0$. The conditional distribution of $X_{1:(k-1)}$ given R_{k-1} and the distribution of R_{k-1} only depend on observations up to time $k-1$ and are given by the recursion. The predictive distribution $p(x_k \,|\, x^{(r_k)})$ only depends on x_k and on the recent observations $x^{(r_k)}$, which enables a recursive algorithm. The calculation of the predictive distribution depends on the chosen model, which will be discussed in Section 4.3.

If we know the change point probability η, the term $\mathbb{P}(R_k = r_k \,|\, R_{k-1} = r_{k-1})$ can be easily determined by

$$\mathbb{P}(R_k = r_k \,|\, R_{k-1} = r_{k-1}) = \begin{cases} 1-\eta & \text{for } r_k = r_{k-1}+1 \\ \eta & \text{for } r_k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

If the change point probability η is unknown, it needs to be estimated as explained in the following (see also Wilson et al., 2010).

4.2.2 Estimation of the change point probability η

If the change point probability η is unknown, we assume a hierarchical Bayesian model in which η is assumed a realization of the random change point probability H. For convenience we use a conjugate prior distribution and thus assume that H is Beta(a_0, b_0)-distributed. Given $\{H = \eta\}$, we assume $Y_1, Y_2, \ldots$ to be independent and Bernoulli(η)-distributed. Due to conjugacy of the distributions, the posterior distribution of H is again a Beta-distribution, cf. Claim 4.2.1 or (Gelman et al., 2013).

Claim 4.2.1. *Let A_k, $k \geq 1$, be Binomial-distributed with parameter (k, H), where H is a random variable, which is Beta-distributed with parameter a_0 and b_0. Then the posterior distribution of H given $\{A_k = a_k\}$ is again a Beta-distribution, i.e.,*

$$H \,|\, \{A_k = a_k\} \sim \mathcal{B}eta(a_0 + a_k, b_0 + k - a_k).$$

Proof. As we choose a $Beta(a_0, b_0)$ prior distribution on H, the density function is

$$\pi_\eta(\eta) = \frac{\Gamma(a_0 + b_0)}{\Gamma(a_0) + \Gamma(b_0)} \eta^{a_0 - 1} (1-\eta)^{b_0 - 1}.$$

Thus we get with Bayes' theorem

$$\begin{aligned} \pi_\eta(\eta \,|\, A_k = a_k) &\sim \pi_\eta(\eta) \cdot \mathbb{P}(A_k = a_k \,|\, H = \eta) \\ &\sim \eta^{a_0 - 1}(1-\eta)^{b_0 - 1} \cdot \eta^{a_k}(1-\eta)^{k - a_k} \\ &\sim \eta^{a_0 + a_k - 1}(1-\eta)^{b_0 + k - a_k}, \end{aligned}$$

which is a $Beta(a_0 + a_k, b_0 + k - a_k)$-distribution. □

In order to determine the posterior run length distribution, cf. Equation (4.1), we sum over all possible numbers of change points A_k up to time k, i.e.,

$$\begin{aligned} \mathbb{P}(R_k = r_k \,|\, X_{0:k} = x_{0:k}) &= \sum_{a_k = 0}^{k} \mathbb{P}(R_k = r_k, A_k = a_k \,|\, X_{0:k} = x_{0:k}) \\ &= \sum_{a_k = 0}^{k} p(x_{0:k} \,|\, R_k = r_k, A_k = a_k) \, \mathbb{P}(R_k = r_k, A_k = a_k) / p(x_{0:k}). \end{aligned}$$

The numerator of every summand can be calculated recursively by

$$\begin{aligned} p(x_{0:k} \,|\, R_k = r_k, &A_k = a_k)\mathbb{P}(R_k = r_k, A_k = a_k) \\ &= \sum_{r_{k-1}} \sum_{a_{k-1}} \mathbb{P}(R_k = r_k, A_k = a_k \,|\, R_{k-1} = r_{k-1}, A_{k-1} = a_{k-1})\, p(x_k \,|\, x^{(r_k)}) \\ &\cdot p(x_{0:k-1} \,|\, R_{k-1} = r_{k-1}, A_{k-1} = a_{k-1})\, \mathbb{P}(R_{k-1} = r_{k-1}, A_{k-1} = a_{k-1}). \end{aligned}$$

The third row is again given by the recursion, and the calculation of the predictive distribution will be discussed in Section 4.3.1.1 (phase) and Section 4.3.2.1 (rate). In order to determine $\mathbb{P}(R_k = r_k, A_k = a_k \,|\, R_{k-1} = r_{k-1}, A_{k-1} = a_{k-1})$, we observe that this term is positive only in two cases; first, if a change point occurs at time k, i.e., if $y_k = 1$ and $\{r_k = 0 \wedge a_k = a_{k-1} + 1\}$, and second, if no change point occurs at time k and the run length increases by one, i.e., $y_k = 0$ and $\{r_k = r_{k-1} + 1 \wedge a_k = a_{k-1}\}$. With Claim 4.2.2 we calculate for the first case, if $0 \leq a_{k-1} \leq k - r_{k-1}$, a probability of

$$\begin{aligned} \mathbb{P}(R_k = 0, A_k = a_{k-1} + 1 \,|\, R_{k-1} = r_{k-1}, A_{k-1} = a_{k-1}) &= \mathbb{P}(Y_k = 1 \,|\, A_{k-1} = a_{k-1}) \\ = \int_0^1 \mathbb{P}(Y_k = 1 \,|\, H = \eta) \cdot \pi_\eta(\eta \,|\, A_{k-1} = a_{k-1})\, d\eta &= \frac{a_0 + a_{k-1}}{a_0 + a_{k-1} + b_0 + b_{k-1}}, \end{aligned} \tag{4.2}$$

where $\pi_\eta(\eta \,|\, A_{k-1} = a_{k-1})$ denotes the posterior $Beta(a_0 + a_{k-1}, b_0 + k - 1 - a_{k-1})$ distribution, and 0 otherwise. For the second case, we obtain the counter-probability

$$\mathbb{P}(R_k = r_{k-1} + 1, A_k = a_{k-1} \,|\, R_{k-1} = r_{k-1}, A_{k-1} = a_{k-1}) = \frac{b_0 + b_{k-1}}{a_0 + a_{k-1} + b_0 + b_{k-1}}.$$

The predictive change point probability (Equation (4.2)) is illustrated in Figure 4.2 D for the path with maximal posterior probability at time $K = 50$.

Claim 4.2.2. *Let $Y_1, \ldots, Y_k$, $k > 1$, be independent and Bernoulli(H)-distributed, where H is a random variable, which is Beta(a_0, b_0)-distributed. Furthermore, let $A_{k-1} := \sum_{i=1}^{k-1} Y_i$ denote the number of successes up to $k - 1$. Then the predictive distribution is given by*

$$\mathbb{P}(Y_k = y_k \,|\, A_{k-1} = a_{k-1}) = \begin{cases} \frac{a_0 + a_{k-1}}{a_0 + a_{k-1} + b_0 + b_{k-1}}, & \textit{if } y_k = 1, \\ \frac{b_0 + b_{k-1}}{a_0 + a_{k-1} + b_0 + b_{k-1}}, & \textit{if } y_k = 0, \\ 0, & \textit{else}, \end{cases}$$

with $b_{k-1} := k - 1 - a_{k-1}$.

Proof. First we note that A_{k-1} is Binomial($k - 1, H$)-distributed and $Y_k \,|\, \{H = \eta\}$ and $A_{k-1} \,|\, \{H = \eta\}$ are independent. Let $\tilde{p}_k := \mathbb{P}(Y_k = 1 \,|\, A_{k-1} = a_{k-1})$ denote the prediction of the change point probability using the information up to time $k - 1$. The law of total probability yields

$$\begin{aligned} \tilde{p}_k &= \int_0^1 \mathbb{P}(Y_k = 1 \,|\, H = \eta, A_{k-1} = a_{k-1}) \cdot \pi_\eta(\eta \,|\, A_{k-1} = a_{k-1}) d\eta \\ &= \int_0^1 \mathbb{P}(Y_k = 1 \,|\, H = \eta) \cdot \pi_\eta(\eta \,|\, A_{k-1} = a_{k-1}) d\eta. \end{aligned}$$

As $\mathbb{P}(Y_k = 1 \mid H = \eta) = \eta$ and using Claim 4.2.1, we get

$$\begin{aligned}\tilde{p}_k &= \int_0^1 \eta \cdot \frac{\Gamma(a_0 + a_{k-1} + b_0 + b_{k-1})}{\Gamma(a_0 + a_{k-1})\Gamma(b_0 + b_{k-1})} \eta^{a_0 + a_{k-1} - 1}(1-\eta)^{b_0 + b_{k-1} - 1} d\eta \\ &= \frac{\Gamma(a_0 + a_{k-1} + b_0 + b_{k-1})}{\Gamma(a_0 + a_{k-1})\Gamma(b_0 + b_{k-1})} \int_0^1 \eta^{a_0 + a_{k-1}}(1-\eta)^{b_0 + b_{k-1} - 1} d\eta \\ &= \frac{\Gamma(a_0 + a_{k-1} + b_0 + b_{k-1})}{\Gamma(a_0 + a_{k-1})\Gamma(b_0 + b_{k-1})} \frac{\Gamma(a_0 + a_{k-1} + 1)\Gamma(b_0 + b_{k-1})}{\Gamma(a_0 + a_{k-1} + 1 + b_0 + b_{k-1})} \\ &= \frac{a_0 + a_{k-1}}{a_0 + a_{k-1} + b_0 + b_{k-1}}\end{aligned}$$

and notice $\mathbb{P}(Y_k = 0 \mid A_{k-1} = a_{k-1}) = 1 - \tilde{p}_k$. □

4.2.3 Algorithmic implementation

Assuming an exponential family sampling model, cf. Section 3.4.1, and a normalizable prior distribution, cf. Section 3.4.3, we can calculate the run length distribution by the following algorithm. Due to computational reasons we do not calculate the value of $p(x_{0:k} \mid R_k = r_k, A_k = a_k)\mathbb{P}(R_k = r_k, A_k = a_k)$ in the recursion, but the unscaled value $\mathbb{P}(R_k = r_k, A_k = a_k \mid X_{0:k} = x_{0:k})$. In the algorithm this term is abbreviated as $\tilde{f}(r_k, a_k, x_{0:k})$, and is passed in the recursive calculation. To compute the run length distribution afterwards, we need to normalize the computation, cf. step 6. Furthermore, let $t(\cdot)$ denote the sufficient statistic of X for θ, cf. Section 3.4.1.

1. Initialize (choose prior parameters of Θ and change point probability H)

$$\mathbb{P}(R_0 = 0) := 1;\ n_0^{(0)} := n_0;\ t_0^{(0)} := t_0;\ a_0 := a_0;\ b_0 := b_0.$$

2. Observe new realization x_k

3. Evaluate Predictive Probability

$$\psi_k^{(j)} := p\left(x_k \mid n_k^{(j)}, t_k^{(j)}\right) \quad \text{for } j = 0, \ldots, k.$$

4. Calculate Growth Probabilities (for $j = 1, \ldots, k$ and $i = a_0, \ldots, k-j$)

$$\begin{aligned}\tilde{f}(r_k = j, a_k = i, x_{0:k}) =& \mathbb{P}(R_k = j, A_k = i \mid R_{k-1} = j-1, A_{k-1} = i) \cdot \psi_k^{(j)} \cdot \\ & \tilde{f}(r_{k-1} = j-1, a_{k-1} = i, x_{0:(k-1)})\end{aligned}$$

5. Calculate change point probabilities (for $i = 1, \ldots, k$)

$$\begin{aligned}\tilde{f}(r_k = 0, a_k = i, x_{0:k}) = \sum_{j=0}^{k-1} & \mathbb{P}(R_k = 0, A_k = i \mid R_{k-1} = j, A_{k-1} = i-1) \cdot \psi_k^{(0)} \cdot \\ & \tilde{f}(r_{k-1} = j, a_{k-1} = i-1, x_{0:(k-1)})\end{aligned}$$

6. Calculate run length distribution (for $j = 0, \dots, k$)

$$\tilde{f}(r_k = j, x_{0:k}) = \sum_{i=1}^{k-j} \tilde{f}(r_k = j, a_k = i, x_{0:k}),$$

   and normalize

$$\mathbb{P}(R_k = j \mid x_{0:k}) = \frac{\tilde{f}(r_k = j, x_{0:k})}{\sum_{i=0}^{k} \tilde{f}(r_k = i, x_{0:k})}.$$

7. Update sufficient statistics (for $j = 1, \dots, k$)

$$\begin{aligned} n_{k+1}^{(0)} &= n_0 \quad \text{and} \quad n_{k+1}^{(j)} = n_k^{(j-1)} + 1 \\ t_{k+1}^{(0)} &= n_0 t_0 \quad \text{and} \quad t_{k+1}^{(j)} = t_k^{(j-1)} + t(x_k). \end{aligned}$$

8. Return to step 2.

4.2.4 Extension: BOCD with online decision

As described above, the BOCD estimates the change points only after observing the complete spike train. However, the brain decides only with minimal delay, and we therefore propose here a modified, 'ad hoc'-algorithm called *BOCD with online decision* which estimates change points directly after a small delay d. Our modified algorithm will help to investigate whether phase parameters can improve this fast decision processes.

The BOCD with online decision is illustrated in Figure 4.3. At cycle $k = 2$ we use only information available up to time $k + d$ and decide in the same way as if the time series ended at time $k + d$. Thus, we detect a change point at cycle k if and only if the most likely run length at time $k + d$ is d, i.e.,

$$\arg\max_{r_{k+d}} \mathbb{P}(R_{k+d} = r_{k+d} \mid X_{0:(k+d)} = x_{0:(k+d)}) = d. \tag{4.3}$$

After cycle $k + d$ it is not possible to estimate a change point at time k. Then we proceed to cycle $k + 1$, taking information up to time $k + 1 + d$, and only consider those possible run lengths that agree with the decision made at cycle k and previous cycles. That means, if a change point was detected at cycle k, only run lengths taking into account that change point are considered. If no change point was detected, only run lengths without such a change point are considered. This considerably reduces the computational effort, but requires that decisions are made almost instantaneously.

To draw on the algorithmic representation of the BOCD in Section 4.2.3 we apply the algorithm to sequences from k to $k + d$ and if Equation (4.3) is fulfilled, we apply the algorithm to sequence $k + 1$ to $k + d + 1$ with prior parameters $a_0 := a_0 + 1$ and $b_0 := b_0$, otherwise with prior parameters $a_0 := a_0$ and $b_0 := b_0 + 1$, cf. '1. Initialize' in the algorithm.

4.3 Application within the change point model

Here we transfer the BOCD, which was described in a general setting using parameter notation θ, to the change point model for rate and phase parameters described in Section 4.1 and Figure

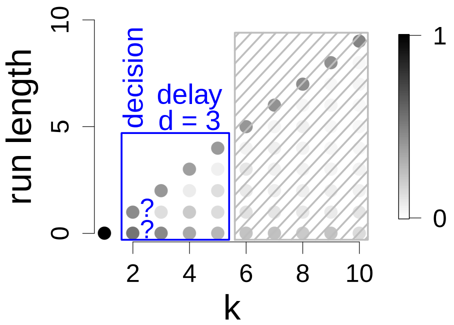


Figure 4.3: Online decision process with delay $d = 3$. The gray scales indicate the probability weight for every possible run length at every time k, i.e., $\mathbb{P}(R_k = r_k \mid X_{0:k} = x_{0:k})$. At time $\tilde{k} = 2$ we use all information up to time $\tilde{k} + d = 5$ and thus decide for the run length which is most likely at time $\tilde{k} + d = 5$. (darkest point for $k = 5$ for a run length of 4, implying $\hat{r}_2 = 1$, i.e., no change is detected at $\tilde{k} = 2$). Without delay, i.e., with $d = 0$, the maximal weight (the darkest point at $k = 2$) would imply an estimated change point at $\tilde{k} = 2$.

4.1. To apply the BOCD we require the predictive distribution $p(x_k \mid x^{(r_k)})$. Here we choose conjugate prior distributions $\pi_\theta(\cdot)$ for θ for convenience, such that the prior and posterior distributions belong to the same distribution family, and even the predictive distribution $p(x_k \mid x^{(r_k)})$ can be determined analytically.
We first recall the parameter setting, then specify the prior and predictive distributions, and finally perform simulations to identify useful parameter choices for the prior distributions. This procedure is done separately for three cases: (1) for the phases, i.e., $\theta = \varphi$, (2) for the rates, i.e., $\theta = \lambda$, (3) for the bivariate case $\theta = (\varphi, \lambda)$. Thus, we always assume that rate and phase parameters change simultaneously, but in our BOCD analysis, we incorporate either only the rate or the phase or the two parameters, and then investigate the advantages of the bivariate change point analysis over univariate analysis.
Note that we now count the cycles from 1 to k as it simplifies the posterior parameter notations. In the notation of the BOCD it was beneficial to count from 0 to k as it simplifies the notation of the recursive update of the change point probability.
In Section 4.3.1 we start with a pure phase analysis and evaluate the performance of the BOCD for different prior parameters and changes in the phase parameter. Thereby we give a short insight to the impact of the prior change point parameters a_0 and b_0, but during the further procedure for generality we decide on an uniform change point prior of $a_0 = b_0 = 1$.
In Section 4.3.2 we focus on a pure rate analysis and investigate the impact of different rate prior parameters. With that we motivate in Section 4.3.3 appropriate and comparable prior parameters for rate and phase in the bivariate analysis and quantify the advantage in the change point detection of the bivariate analysis compared to a pure rate analysis applying the BOCD and the BOCD with online decision.

4.3.1 Pure phase analysis

In the following sections we analyze the ability of the phase by its own detect changes in the timing of the spikes. First we assume that exactly one spike in each oscillation cycle occurs, i.e., $n_k = 1 \ \forall k \geq 1$. To apply the BOCD we need to calculate the predictive distribution of

$X_{k+1} \,|\, X_{1:k}$. Therefore, we determine in Section 4.3.1.1 a conjugate prior distribution and its posterior update process and calculate finally the predictive distribution, cf. Gelman et al. (2013). Afterwards we apply the BOCD to the setting of one spike per oscillation cycle and consider the two cases no and one change point, see Section 4.3.1.2. Thereby we give a deeper understanding of the impact of the chosen prior change point parameters a_0, b_0 and prior parameters of the phase and how they influence the change point detection.

In Section 4.3.1.3 we extend the results of Section 4.3.1.1 to an arbitrary spike number n_k in each oscillation cycle and state how the predictive distribution can be calculated if we observe multiple random number of spikes in each cycle. With that we are able to search for change points in the phase in our change point model and analyze the impact of the spike rate in the change detection based on a pure phase analysis.

4.3.1.1 One spike

We assume we observe one spike in each oscillation cycle, i.e., $n_k = 1 \,\forall\, k \geq 1$. Recall that in our change point model in each cycle k, given the phase parameter $\varphi := \varphi_{A_k}$, we observe $X_k := X_1^{(k)} \sim \mathcal{N}(\varphi, \sigma^2)$. Our aim is to correctly detect the positions, where a change in the mean parameter of the normal distribution occurs. To use the BOCD to detect change point in the phase, we need to determine the predictive distribution. Therefore let us first specify a conjugate prior distribution of Φ and determine the posterior distribution of $\Phi \,|\, X_{1:k}$.

Lemma 4.3.1. *Let $\Phi \sim \mathcal{N}(\mu_0, \tau_0^2)$ be the prior distribution and $X_1, \ldots, X_k$ a sequence of i.i.d. random variable with $X_1 \,|\, \{\Phi = \varphi\} \sim \mathcal{N}(\varphi, \sigma^2)$, where σ^2 is known. Then the posterior distribution $\Phi \,|\, \{X_{1:k} = x_{1:k}\}$ is again a normal distribution, i.e.,*

$$\Phi \,|\, \{X_{1:k} = x_{1:k}\} \sim \mathcal{N}(\mu_k, \tau_k^2),$$

where

$$\tau_k^2 := \frac{1}{\frac{1}{\tau_0^2} + \frac{k}{\sigma^2}} \quad \textit{and} \quad \mu_k := \frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum_{i=1}^k x_i}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{k}{\sigma^2}}.$$

Proof. Let $\pi_\varphi(\cdot)$ denote the prior distribution of Φ and $p(\cdot)$ the distribution of $X_{1:k}$ and $\phi_{\varphi,\sigma^2}(\cdot)$ the density of the normal distribution with mean φ and variance σ^2. As $X_1, \ldots, X_k$ are conditional independent, we obtain $p(x_{1:k} \,|\, \Phi = \varphi) = \prod_{i=1}^k \phi_{\varphi,\sigma^2}(x_i)$.

As $\bar{X} := \sum_{i=1}^k X_i/k$ is sufficient for φ, we will reduce the problem to the univariate case by reverting to the sample mean $\bar{X}$, which notoriously has distribution $\bar{X} \,|\, \{\Phi = \varphi\} \sim \mathcal{N}(\varphi, \sigma^2/k)$, i.e.,

$$\begin{aligned}
p(x_1, \ldots, x_k \,|\, \Phi = \varphi) &\sim \exp\left(-1/(2\sigma^2) \sum_{i=1}^k (x_i - \varphi)^2\right) \\
&\sim \exp\left(-1/(2\sigma^2) \left(\sum x_i^2 - 2\varphi \sum x_i + k\varphi^2\right)\right) \\
&\sim \exp\left(-k/(2\sigma^2)\left(-2\varphi\bar{x} + \varphi^2\right)\right) \\
&\sim \exp\left(-k/(2\sigma^2)\left(\bar{x} - \varphi\right)^2\right),
\end{aligned}$$

which is a $\mathcal{N}(\varphi, \sigma^2/k)$-distribution. Using Bayes Rule yields

$$\begin{aligned}
\pi_\varphi(\varphi \mid X_{1:k} = x_{1:k}) &\sim \pi_\varphi(\varphi) \cdot p(x_{1:k} \mid \Phi = \varphi) \\
&\sim \exp(-1/(2\tau_0^2)(\varphi - \varphi_0)^2) \cdot \exp\left(-k/(2\sigma^2)\,(\bar{x} - \varphi)^2\right) \\
&\sim \exp\left(\frac{1}{2}\left(\frac{1}{\tau_0^2} + \frac{k}{\sigma^2}\right)\left(\frac{\left(\frac{\varphi}{\tau_0} - \frac{\mu_0}{\tau_0}\right)^2 + \frac{k}{\sigma^2}(\varphi - \bar{x})^2}{1/\tau_0^2 + n\sigma^2}\right)\right) \\
&\sim \exp\left(\frac{1}{2\tau_k^2}\left(\frac{\frac{\varphi^2}{\tau_0^2} - 2\varphi\frac{\varphi_0}{\tau_0^2} + \frac{\mu_0^2}{\tau_0^2} + \varphi\frac{k}{\sigma^2} - 2\varphi\frac{k}{\sigma^2}\bar{x} + \frac{k}{\sigma^2}\bar{x}^2}{1/\tau_0^2 + k\sigma^2}\right)\right) \\
&\sim \exp\left(\frac{1}{2\tau_k^2}\left(\frac{\varphi^2\left(\frac{1}{\tau_0^2} + \frac{k}{\sigma^2}\right) - 2\varphi\left(\frac{\mu_0^2}{\tau_0^2} + k\frac{\bar{x}}{\sigma^2}\right) + \frac{\mu_0^2}{\tau_0^2} + \frac{k}{\sigma^2}\bar{x}^2}{1/\tau_0^2 + k\sigma^2}\right)\right) \\
&\sim \exp\left(\frac{1}{2\tau_k^2}\left(\varphi - 2\varphi\frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum_{i=1}^k x_i}{\sigma^2}}{1/\tau_0^2 + k/\sigma^2}\right)\right) \\
&\sim \exp\left(\frac{1}{2\tau_k^2}\,(\varphi - \mu_k)^2\right).
\end{aligned}$$

□

Remark 1: In Algorithm 4.2.3 the sufficient statistic can be updated recursively by

$$\begin{aligned}
\mu_{k+1}^{(0)} = \mu_0 \quad \text{and} \quad \mu_{k+1}^{(j)} &= \frac{\frac{\mu_k^{(j-1)}}{\tau_k^{(j-1)}} + \frac{x_k}{\sigma^2}}{1/\tau_k^{(j-1)} + 1/\sigma^2} \\
\tau_{k+1}^{(0)} = \tau_0^2 \quad \text{and} \quad \tau_{k+1}^{(j)} &= \frac{1}{1/\tau_k^{(j-1)} + 1/\sigma^2},
\end{aligned}$$

as according to Lemma 4.3.1

$$\begin{aligned}
\tau_{k+1}^2 &= \frac{1}{1/\tau_0^2 + (k+1)\sigma^2} \\
&= \frac{1}{1/\tau_0^2 + k/\sigma^2 + 1/\sigma^2} = \frac{1}{\tau_k^2 + 1/\sigma^2}
\end{aligned}$$

and

$$\begin{aligned}
\mu_{k+1} &= \frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum_{i=1}^{k+1} x_i}{\sigma^2}}{1/\tau_0^2 + (k+1)/\sigma^2} \\
&= \frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum_{i=1}^k x_i}{\sigma^2} + \frac{x_{k+1}}{\sigma^2}}{1/\tau_0^2 + k/\sigma^2 + 1/\sigma^2} = \frac{\frac{\mu_k}{\tau_k^2} + \frac{x_{k+1}}{\sigma^2}}{1/\tau_k^2 + 1/\sigma^2}.
\end{aligned}$$

Remark 2: The posterior parameters τ_k^2 and μ_k have a nice interpretation:
The equation

$$\frac{1}{\tau_k^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2/k}$$

can be interpreted as

$$\textit{posterior precision} = \textit{prior precision} + \textit{sampling precision}.$$

The location parameter μ_k is just the weighted average of the prior mean parameter and the sample mean.
The update process of the posterior parameters for $k = 1, 5, 10$ observations is illustrated in Figure 4.4 A. The corresponding update of the predictive distribution is in Figure 4.4 B and is calculated according to Lemma 4.3.2, which says: The predictive distribution of a normal distribution on basis of k observations is again a normal distribution with posterior expectation μ_k and variance the uncertainty about Φ, which is τ_k^2, plus the uncertainty about a random realization, which is σ^2.

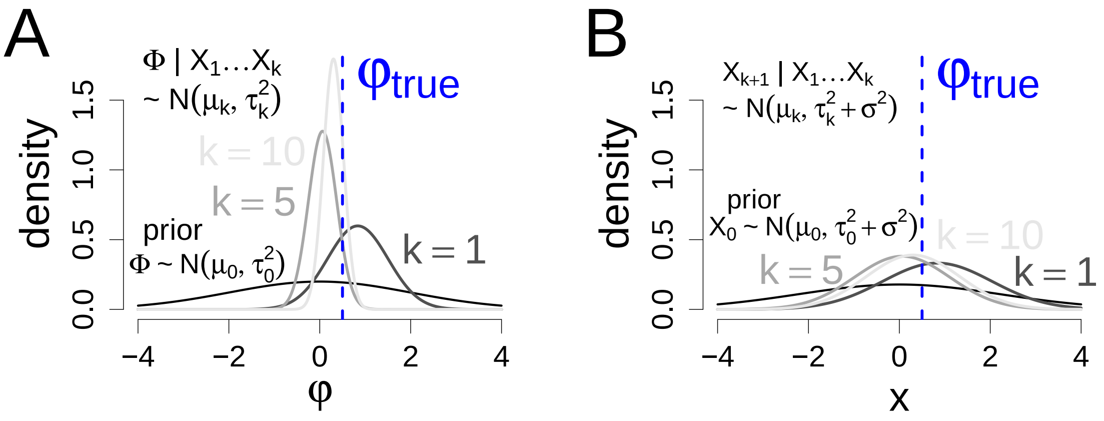


Figure 4.4: A. Update process of the posterior parameters μ_k and τ_k of the phase variable Φ for $k = 1, 5, 10$ observations. The true (unknown) phase parameter is $\varphi_{true} = 0.5$. The known variance is $\sigma^2 = 1$. As prior parameters we choose $\mu_0 = 0$ and $\tau_0^2 = 4$. Thus starting with a broad $\mathcal{N}(0, 4)$-prior-distribution, we get closer to the true phase parameter φ_{true} by updating the posterior parameters μ_k and τ_k according to Lemma 4.3.1. B. The predictive distribution of $X_{k+1} \,|\, \{X_{1:k}\}$ for the same parameters and observations as in A. Thus starting with $\mathcal{N}(0, 5)$-prior-predictive-distribution the posterior predictive distribution converges to a $\mathcal{N}(0.5, 1)$-distribution for a increasing number of observations k.

Lemma 4.3.2. *Let $\Phi \sim \mathcal{N}(\mu_0, \tau_0^2)$ and $X_1, \ldots, X_k$ be a sequence of i.i.d. random variable with $X_1 \,|\, \{\Phi = \varphi\} \sim \mathcal{N}(\varphi, \sigma^2)$, where σ^2 is known. Then the predictive distribution $X_{k+1} \,|\, \{X_{1:k} = x_{1:k}\}$ is again normally distributed, i.e.,*

$$X_{k+1} \,|\, \{X_{1:k} = x_{1:k}\} \sim \mathcal{N}(\mu_k, \tau_k^2 + \sigma^2),$$

where μ_k and τ_k are defined as in Lemma 4.3.1.

Proof. As X_{k+1} arises through a hierarchical model, where we first draw a normally distributed variable Φ and given $\{\Phi = \varphi\}$ we choose X_{k+1} according to a $\mathcal{N}(\varphi, \sigma)$-distribution, we know $X_{k+1} \sim \mathcal{N}(\mu_0, \tau_0^2 + \sigma^2)$ and can fragment

$$X_{k+1} = \Phi + Z,$$

where $Z \sim \mathcal{N}(0, \sigma^2)$ and independent of Φ. Out of Lemma 4.3.1 we know $\Phi \,|\, \{X_{1:k} = x_{1:k}\} \sim \mathcal{N}(\mu_k, \tau_k^2)$. Furthermore, $Z \,|\, \{X_{1:k} = x_{1:k}\} \sim Z \sim \mathcal{N}(0, \sigma^2)$, so $X_{k+1} \,|\, \{X_{1:k} = x_{1:k}\}$ is also normally distributed. In the following we determine the parameters of the normal distribution:

$$\begin{aligned}\mathbb{E}[X_{k+1} \,|\, X_{1:k} = x_{1:k}] &= \mathbb{E}[\Phi + Z \,|\, X_{1:k} = x_{1:k}] \\ &= \mathbb{E}[\Phi \,|\, X_{1:k} = x_{1:k}] + \underbrace{\mathbb{E}[Z \,|\, X_{1:k} = x_{1:k}]}_{=0} \\ &= \mu_k\end{aligned}$$

and

$$\begin{aligned}\mathbb{V}ar[X_{k+1} \,|\, X_{1:k} = x_{1:k}] &= \mathbb{V}ar[\Phi \,|\, X_{1:k} = x_{1:k}] + \mathbb{V}ar[Z \,|\, X_{1:k} = x_{1:k}] \\ &= \tau_k^2 + \sigma^2.\end{aligned}$$

□

Remark 4.3.3. *The variance equation also follows directly from the law of total variance, which says for $X \sim \mathcal{N}(\Phi, \sigma^2)$*

$$\begin{aligned}\mathbb{V}ar[X] &= \mathbb{V}ar\left[\mathbb{E}_\Phi[X]\right] + \mathbb{E}\left[\mathbb{V}ar_\Phi[X]\right] \\ &= \mathbb{V}ar\left[\Phi\right] + \sigma^2.\end{aligned}$$

4.3.1.2 Application of BOCD (1 Spike)

With the preceding section we know the predictive distribution and can use the BOCD, cf. Section 4.2.3 to search for change points. Therefore, we consider the two cases no change point or one change point occur and analyze the impact of the number of observed cycles, the prior change point parameters a_0 and b_0 and the prior precision τ_0^2.

No change point: Here we study the scenario that no change point occurs and we want to test how the prior change point parameters a_0 and b_0 and the number of cycles K influence the ability of the BOCD to correctly detect no change point. So we just observe $\mathcal{N}(0, 1)$-random-variables. As prior location parameter we choose $\mu_0 = 0$ and as prior precision $\tau_0 = 1$. In Figure 4.5 A we observed $K = 100$ random realization and applied the BOCD with $a_0 = 1$ and different b_0.

If we do not have any special information about the change point frequency, we may choose $a_0 = 1$ and $b_0 = 1$, which is the uniform distribution. If we expect rather less change points, we would choose $b_0 > 1$, since the expectation of a Beta(a_0, b_0)-distribution is $a_0/(a_0 + b_0)$. In some simulations starting with an uniform change point prior results in many detected change points. Choosing $b_0 < 1$ increases the number of wrong detection. Increasing b_0 up to $K/2$ results in almost no false detections, cf. Figure 4.5 A.

Increasing the number of cycles up to $K = 500$ provides enough information to remove almost

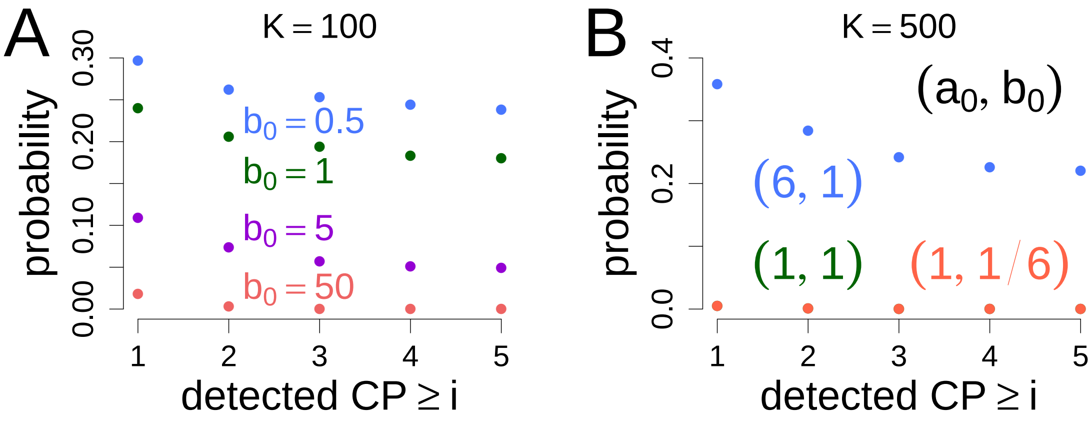


Figure 4.5: We consider K independent observations from a standard normal distribution and analyze the number of detected change points applying the BOCD with $\mu_0 = 0$ and $\tau_0 = 1$. A. Here we consider $K = 100$ observations and choose $a_0 = 1$. A strong concern of a small change point probability ($b_0 \gg 1$) is needed for a small number of detected change points. B. For $K = 500$ observations even an excessive change point prior yields almost no change point detection (Beta$(1, 1/6)$), thus only some chance for a small change point probability is needed (the density of a Beta$(6, 1$ distribution) equals zero at zero).

all falsely detected change points, even for a change point prior with a high expected number of changes, cf. Figure 4.5 B (red dots). But even in case of many cycles, there is some caution needed in the choice of the change point prior. For example, if there is almost no chance for a small change point probability in the prior distribution (blue dots), even a large time horizon can not fix the overestimation, for explanation check Figure 4.6.

One change point: Here we study the scenario that exactly one change point occurs at $K/2$ for $K = 100$ observations. Again a higher number of cycles improves the detection ability. We observe $K/2$ $\mathcal{N}(0, 1)$-random-variables and $K/2$ $\mathcal{N}(\varphi_c, 1)$-random-variables. Thereby we analyze different changes in the location parameter φ_c. Once more we choose $\mu_0 = 0$. But this time we analyze the impact of prior precision τ_0. In Figure 4.7 A. we choose $\tau_0 = 1$.
A Beta$(1, K/2)$-prior-change point distribution detect changes in φ starting from $\varphi_c = 0.5$ and almost make no false detections. If we start with a uniform change point prior distribution ($a_0 = b_0 = 1$), there are many falsely detected change points. But if we decrease our prior precision and choose $\tau_0 = 2$, also a uniform change point prior yields a few false detections, cf. Figure 4.7 B.
The majority of falsely detected change points in case of $\tau_0 = 1$ and a uniform change point prior results from simulations, where the algorithm detect changes at every fourth observation, cf. Figure 4.8 A. For a prior precision $\tau_0 = 2$ the BOCD detects for almost all simulations less than 6 change points, cf. Figure 4.8 B. Thus the divergence (K not to large) in some simulations results from a bad prior choice, which can be seen heuristically in the following way:
In general a high prior precision (τ_0 is small) causes an overestimation of the change point number (see also subsection 4.3.1.3), if we start in a broad change point prior distribution

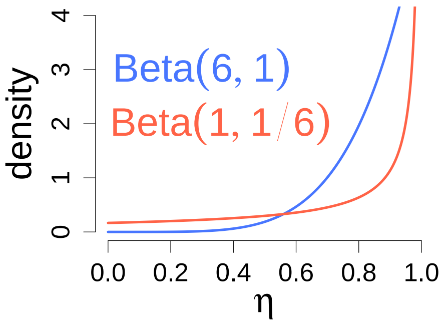


Figure 4.6: Comparison of a Beta$(6, 1)$- and Beta$(1, 1/6)$-density.

Both change point priors Beta$(6, 1)$ and Beta$(1, 1/6)$ have the same expectation $6/7$, but only Beta$(1, 1/6)$ results in almost no detected change point for a high number of $K = 500$ cycles. This is due to the structure of the densities: Beta$(6, 1)$ is more flat and equals zero at 0. Thus small change point probabilities are not only unlikely, but almost impossible. This results in a structural overestimation of the change point number. In case of a Beta$(1, 1/6)$, small values for p are unlikely, but not impossible, thus $K = 500$ cycles can fix the misinformation.

(e.g. uniform distribution) and consider a limited time horizon, as:
Let us assume we observed k realizations, where we decided, not to detect a change point. Then at time $k + 1$ we balance the two possibilities

$$\begin{aligned} &1) \quad X_{k+1} \sim \mathcal{N}(\mu_0, \tau_0^2 + \sigma^2) \\ &2) \quad X_{k+1} \sim \mathcal{N}(\mu_k, \tau_k^2 + \sigma^2). \end{aligned}$$

If $\mathbb{E}[X_i] = \mu_0$ (as in the simulations), we know according to the law of large numbers:

$$\mu_k = \frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum X_i}{\sigma^2}}{1/\tau_0^2 + k/\sigma^2} \xrightarrow{k\to\infty} \mu_0.$$

Furthermore we have $\tau_k < \tau_0 \;\forall\, k \geq 1$, since

$$\tau_k^2 = \frac{1}{1/\tau_0^2 + k/\sigma^2} = \tau_0^2 \left(\frac{1}{1 + k\frac{\tau_0^2}{\sigma^2}} \right).$$

Additionally the larger τ_0/σ is, the smaller τ_k/τ_0 gets and vice versa, i.e.,

$$\frac{\tau_0}{\sigma} \nearrow \Longrightarrow \frac{\tau_k}{\tau_0} \searrow .$$

Thus if we choose a higher prior precision, which is equivalent to decrease τ_0/σ, the difference of the two possibilities $\mathcal{N}(\mu_0, \tau_0^2 + \sigma^2)$ or $\mathcal{N}(\mu_k, \tau_k^2 + \sigma^2)$ decreases. Thereby the high change point prior influences the change point detection essentially.

4.3.1.3 Arbitrary spike number

In the following section we extend the results of Section 4.3.1.1 to an arbitrary spike number. Recall that in our change point model, in each cycle k, given the phase parameter $\varphi := \varphi_{A_k}$ and the spike number n_k, we observe independent random variables $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ with $X_j^{(k)} \sim \mathcal{N}(\varphi, \sigma^2)$. As the mean spike time $\bar{X}_k := 1/n_k \sum_{j=1}^{n_k} X_j^{(k)}$ is sufficient for φ, we

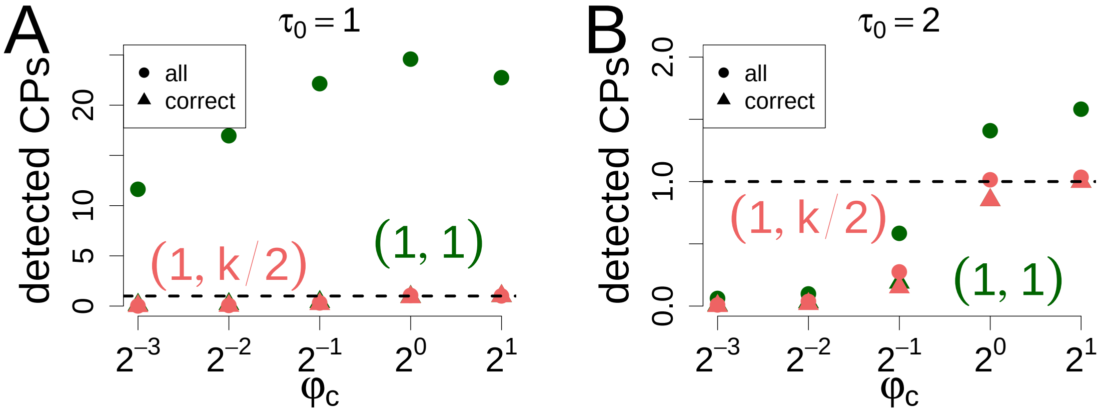


Figure 4.7: We consider $K = 100$ independent observations, where the first $K/2$ follow a $\mathcal{N}(0, 1)$-distribution and the other half a $\mathcal{N}(\varphi_c, 1)$-distribution. Dependent on the change in the location parameter the overall number and the correct number of change points is shown. A change point is correct detected, if the distance to the true change point is less than 3. A. We choose a prior precision $\tau_0 = 1$. In that case we detect a high number of false change points with an uniform change point prior. B. We choose a prior precision $\tau_0 = 2$. Here we obtain plausible results for both change point priors.

consider only $\bar{X}_k \sim \mathcal{N}(\varphi, \sigma^2/n_k)$ for the posterior distribution if we observe at least one spike. Empty cycles with $n_k = 0$ are skipped.
The conjugate prior distribution of the phase parameter is then a normal distribution, i.e., $\Phi_0 \sim \mathcal{N}(\mu_0, \tau_0^2)$, and the posterior distribution $\Phi_0 \,|\, \{\bar{X}_{1:k} = \bar{x}_{1:k}\}$ after $k+1$ cycles without a change point is, cf. Lemma 4.3.1,

$$\Phi_0 \,|\, \{\bar{X}_{1:k} = \bar{x}_{1:k}\} \sim \mathcal{N}(\mu_k, \tau_k^2), \quad \text{where} \quad \tau_k^2 := \frac{1}{\frac{1}{\tau_0^2} + \frac{\sum_{i=1}^k n_i}{\sigma^2}} \quad \text{and} \quad \mu_k := \frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum_{i=1}^k n_i \bar{x}_i}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{\sum_{i=1}^k n_i}{\sigma^2}}.$$

To calculate the predictive distribution for cycle $k+1$, we update the distribution of φ successively in the cycle, i.e.,

$$p\left(X_{1:n_{k+1}}^{(k+1)} = x_{1:n_{k+1}}^{(k+1)} \,|\, \bar{X}_{1:k} = \bar{x}_{1:k}\right) = p\left(X_1^{(k+1)} \,|\, \bar{X}_{1:k} = \bar{x}_{1:k}\right) \cdot \ldots \cdot p\left(X_{n_{k+1}}^{(k+1)} \,|\, X_{1:n_{k+1}-1}^{(k+1)} = x_{1:n_{k+1}-1}^{(k+1)}, \bar{X}_{1:k} = \bar{x}_{1:k}\right).$$

On the right-hand side, the predictive distribution of $X_{i+1}^{(k+1)} \,|\, \{X_{1:i}^{(k+1)} = x_{1:i}^{(k+1)}, \bar{X}_{1:k} = \bar{x}_{1:k}\}$, $i = 0, \ldots, k$, is a normal distribution with mean μ_{ki} and variance $\tau_{ki}^2 + \sigma^2$ given by, cf. Lemma 4.3.2,

$$\tau_{ki}^2 := \frac{1}{\frac{1}{\tau_k^2} + \frac{i}{\sigma^2}} \quad \text{and} \quad \mu_{ki} := \frac{\frac{\mu_k}{\tau_k^2} + \frac{\sum_{j=1}^i x_j^{(k+1)}}{\sigma^2}}{\frac{1}{\tau_k^2} + \frac{i}{\sigma^2}}. \tag{4.4}$$

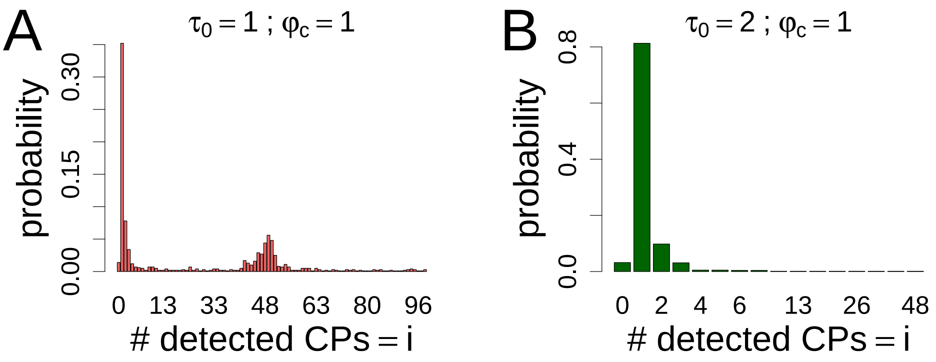


Figure 4.8: We consider $K = 100$ independent observations, where the first $K/2$ follow a $\mathcal{N}(0, 1)$-distribution and the other half a $\mathcal{N}(\varphi_c, 1)$-distribution, with $\varphi_c = 1$. In 10000 simulations we choose a uniform change point prior ($a_0 = b_0 = 1$) and determine the number of detected change points for two different priors. A. With a prior precision $\tau_0 = 1$ we detect a huge number of change points in several simulations. B. Reducing the prior precision to $\tau_0 = 2$ yields almost no falsely detected change points.

The successive update to calculate the predictive distribution is crucially, and we can not combine the observations in one cycle to the mean spike time and use $p(\bar{x}_{k+1} \mid \bar{X}_{1:k} = \bar{x}_{1:k})$ in the BOCD. We will discuss the difference and its impact on the change point detection in Section 4.3.3.4.

4.3.1.4 Application of BOCD (arbitrary spike number)

Now we are able to apply the BOCD to detect changes in the phase parameter if we have an arbitrary (fixed or random) spike number per oscillation cycle.
First we consider a no change point scenario and compare the robustness of the phase for 1, 2 or 4 spikes per oscillation cycle. Thereby we vary the prior precision τ_0 of the phase and examine how a miss-specification in the prior belief μ_0 influence the change point detection.
Second we consider an one change point scenario and analyze how the change detection with a random number of spikes ($N \sim Pois(\lambda)$) differs from a fixed number of spikes ($N = \lambda$, $\lambda \in \mathbb{N}$).
Third we consider various sizes of the change in an one change point scenario and compare one spike with two spikes per oscillation cycle, but we consider twice as much cycles in the one spike setting.
In each scenario we start with an uniform change point prior ($a_0 = b_0 = 1$).

No change point We consider various numbers of cycles $K = 10, 20, \ldots, 200$ containing no change point and three scenarios of $n_k = 1, 2, 4$ spikes per oscillation cycle ($\forall k$), i.e., for $k = 1, \ldots, K$ we observe

$$X_1^{(k)}, \ldots, X_{n_k}^{(k)}, \text{ where } X_j^{(k)} \sim \mathcal{N}(0, 1) \ \forall k, j \text{ and independent,}$$

see Figure 4.9. Thereby we examine three types of prior belief:
In Figure 4.9 A and B we have a correct prior expectation $\mu_0 = 0$, but in Figure 4.9 A we have a very strong opinion $\tau_0 = 0.5$ and in Figure 4.9 B we choose a wide prior distribution

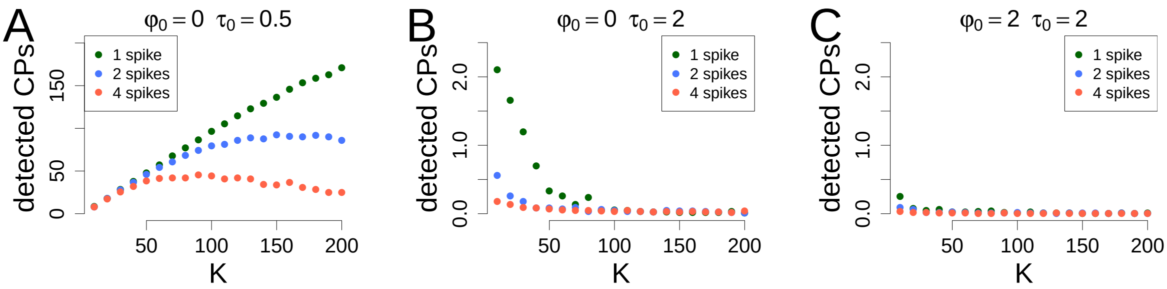


Figure 4.9: No change point scenario: We observe a sequence of $K = 10, 20, \ldots, 200$ oscillation cycles, where in each cycle the spike times are independent and standard normally distributed. We consider 1 (green), 2 (blue) or 4 (red) spikes per oscillation cycle and determine the number of detected change points. A. Precise prior belief: We have prior expectation $\mu_0 = 0$ and precision $\tau_0 = 0.5$. B. Imprecise prior belief: We have prior expectation $\mu_0 = 0$ and precision $\tau_0 = 2$. C. Miss-specified prior belief: We have prior expectation $\mu_0 = 2$ and precision $\tau_0 = 2$.

with $\tau_0 = 2$. If we have a very precise prior opinion we detect many change points and need a long time horizon (K large) to reduce the number of detected change points. This issue is connected to the uniform change point prior, which is to broad to start which such precise prior information about Φ, cf. Section 4.3.1.2. However, the higher the number of spikes per oscillation cycle, the lower the number of detected change points and the faster an increased number of cycles can fix the problem of the change point prior. A broad prior belief (Figure 4.9 B) results in a robust change detection and only for one spike per oscillation cycle (green dots) we observe a considerable number of detected change points, which reduces to zero almost completely for $K > 80$.

In Figure 4.9 C we start with a wrong prior expectation $\mu_0 = 2$. The result is a very robust change detection (for all number of spikes and observed oscillation cycles), which might seem surprising at first sight. But the miss specification leads to low prior density for the relevant parameter range and already after one observation the posterior distribution fits much better than the prior distribution, which compensates the difficulty of an uniform change point prior directly. However, it is difficult to detect a small change, for example $\varphi : 0 \to 0.5$, if we have such a wrong prior expectation, as we would again prefer the posterior distribution, due to the low prior density.

One change point In Figure 4.10 A and B we compare a random spike number ($N \sim Pois(\lambda)$, Figure 4.10 A) with a fixed spike number ($N = \lambda$, $\lambda \in \mathbb{N}$ Figure 4.10 B). Therefore we consider various numbers of cycles $K = 10, 20, \ldots, 100$ with exactly one change point in the mid at $K/2$, i.e., for all $k = 1, \ldots, K/2$ we observe given $\{N_k = n_k\}$

$$X_1^{(k)}, \ldots, X_{n_k}^{(k)}, \text{ where } \forall\ j = 1, \ldots, n_k\ X_j^{(k)} \text{ are i.i.d. with } X_1^{(k)} \sim \mathcal{N}(0, 1)$$

and for all $k = K/2 + 1, \ldots, K$ given $\{N_k = n_k\}$

$$X_1^{(k)}, \ldots, X_{n_k}^{(k)}, \text{ where } \forall\ j = 1, \ldots, n_k\ X_j^{(k)} \text{ are i.i.d. with } X_1^{(k)} \sim \mathcal{N}(0.5, 1).$$

Whether the spike number is fixed or random, the performance of the change detection is quite similar, i.e.: The number of correctly detected change point (circles) and the number of all

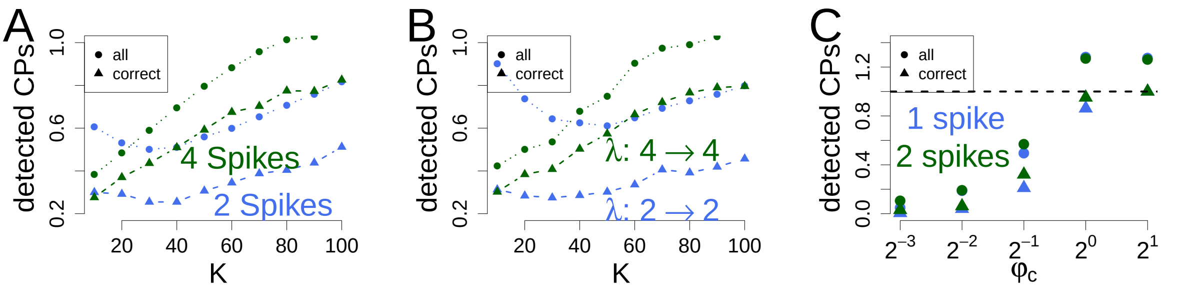


Figure 4.10: One change point scenario with prior parameters $\mu_0 = 0$ and $\tau_0 = 2$. In the first $K/2$ cycles we observe standard normally distributed spike times. We distinguish the number of all detected change points (dots) and only the number of correct detected change points (triangles). A. and B. Exactly one change point at cycle $K/2$ for $K = 10, 20, \ldots, 100$ cycles. The phase jumps from 0 to 1/2. A. Fixed spike number 2 (blue) and 4 (green) per oscillation cycle. B. Poisson(λ) distributed spike number $\lambda = 2$ (blue) and $\lambda = 4$ (green). C. We consider $K = 50$ cycles, green, ($K = 100$, blue) with 2 spikes (1 spike) each cycle and exactly one change point at cycle $K/2$, where the phase jumps from 0 to $\varphi_c = 2^{-3}, 2^{-2}, \ldots, 2^{1}$.

detected change points, falsely plus correctly, (triangles) are almost identical. At the very most in case of a spike rate $\lambda = 2$ (blue) and few cycles ($K \leq 30$), a random spike number results in a higher number of falsely detected change points (blue dots): A Poisson(2)-distribution creates with about 27% one spike in an oscillation cycle, which significantly increases the variance of the mean spike time in the first cycles.

In Figure 4.10 C we question whether there is an advantage to observe two spikes in each cycle and $K = 50$ cycles (green) or instead observe only one spike in each cycle, but $2K = 100$ cycles (blue). Therefore, we use a wide prior distribution $\tau_0 = 2$ with a prior expectation $\mu_0 = 0$ and consider changes in the mid at $K/2$ of $0 \to \varphi_c$, with $\varphi_c = 2^{-3}, 2^{-2}, \ldots, 2^{1}$. Both settings perform almost identical, however 2 spikes have a small advantage for a change of size 2^{-1} and 2^{0}. Maybe one would have expected that, as in the case of 2 spikes we know for certain, that these two spikes have the same phase, on the contrary in case of 1 spike, two adjacent spikes can have a different phase parameter, because they belong to different cycles. Thus observing more spikes per cycle should be an advantage, but is rather decisive for few cycles.

4.3.2 Pure rate analysis

In the following section we analyze the ability of the rate to detect changes in the number of spikes. In each cycle we observe a Poisson distributed number of spikes, where the parameter of the Poisson distribution is random and can change between two cycles. Our aim is to detect such changes in the parameter.

To be able to apply the BOCD we need access to the predictive distribution $N_{k+1} \mid N_{1:k}$. Therefore, in Section 4.3.2.1 we specify a conjugate prior distribution for Λ, determine the posterior update process of $\Lambda \mid N_{1:k}$ and finally calculate the predictive distribution, cf. Gelman et al. (2013).

In Section 4.3.2.2 we first apply the BOCD to a no change point scenario and compare different rate parameters with various prior distributions. Afterwards we consider those prior distributions in an one change point scenario with two different rate changes.

4.3.2.1 Predictive distribution

First we give a short overview of the prior, posterior and predictive distribution for the rate parameter. In each cycle k, given the realization of the rate parameter $\lambda := \lambda_{A_k}$, we observe a Poisson distributed number of spikes $N_k \sim \mathit{Pois}(\lambda)$. We choose the conjugate distribution $\Lambda_0 \sim \mathcal{G}\mathit{amma}(\alpha_0, \beta_0)$ as a prior distribution. After $k+1$ cycles without change point, the posterior distribution is again a Gamma distribution, cf. Lemma 4.3.4, i.e.,

$$\Lambda_0 \,|\, \{N_{1:k} = n_{1:k}\} \sim \mathcal{G}\mathit{amma}(\alpha_k, \beta_k),$$

with $\alpha_k := \alpha_0 + \sum_{i=1}^k n_i$ and $\beta_k := \beta_0 + k$. The predictive distribution of $N_{k+1} \,|\, \{N_{1:k} = n_{1:k}\}$ is a negative binomial distribution, cf. Lemma 4.3.5, i.e.,

$$N_{k+1} \,|\, \{N_{1:k} = n_{1:k}\} \sim \mathcal{NB}\left(\alpha_k, \frac{\beta_k}{\beta_k + 1}\right). \tag{4.5}$$

In Figure 4.11 an illustration of the posterior update process and of the predictive distribution can be found.

Lemma 4.3.4. *Let $\Lambda \sim \mathcal{G}\mathit{amma}(\alpha_0, \beta_0)$ be the prior distribution and $N_1, \dots, N_k$ a sequence of i.i.d. random variable with $N_1 \,|\, \{\Lambda = \lambda\} \sim \mathit{Pois}(\lambda)$. Then the posterior distribution $\Lambda \,|\, \{N_{1:k} = n_{1:k}\}$ is again a Gamma distribution, i.e.,*

$$\Lambda \,|\, \{N_{1:k} = n_{1:k}\} \sim \mathcal{G}\mathit{amma}(\alpha_k, \beta_k),$$

where

$$\alpha_k := \alpha_0 + \sum_{i=1}^k n_i \quad \textit{and} \quad \beta_k := \beta_0 + k.$$

Proof. Let $\pi_\lambda(\cdot)$ denote the prior distribution of Λ, i.e.,

$$\pi_\lambda(\lambda) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda} \quad (\lambda > 0),$$

where $\Gamma(\cdot)$ denotes the gamma function (see Definition B.1). As $N_1, \dots, N_k$ are conditional independent, i.e., $\mathbb{P}(N_{1:k} = n_{1:k} \,|\, \Lambda = \lambda) = \prod_{i=1}^k \mathbb{P}(N_i = n_i \,|\, \Lambda = \lambda)$, Bayes Rule yields

$$\begin{aligned}\pi_\lambda(\lambda \,|\, N_{1:k} = n_{1:k}) &\sim \pi_\lambda(\lambda) \cdot \mathbb{P}(N_{1:k} = n_{1:k} \,|\, \Lambda = \lambda) \\ &\sim \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda} \cdot \frac{\lambda^{\sum_{i=1}^k n_i}}{\prod_{i=1}^k n_i!} e^{-\lambda k} \\ &\sim \lambda^{\alpha_0 + \sum_{i=1}^k n_i - 1} e^{-(\beta_0 + k)\lambda},\end{aligned}$$

which is a Gamma distribution with parameters $\alpha_0 + \sum_{i=1}^k n_i$ and $\beta_0 + k$. □

Remark 1: In Algorithm 4.2.3 the sufficient statistic can be updated recursively by

$$\alpha_{k+1}^{(0)} = \alpha_0 \quad \text{and} \quad \alpha_{k+1}^{(j)} = \alpha_k^{j-1} + n_k$$
$$\beta_{k+1}^{(0)} = \beta_0 \quad \text{and} \quad \beta_{k+1}^{(j)} = \beta_k^{j-1} + 1.$$

Remark 2: As the mean of the gamma distribution is (Remark B.5)

$$\mathbb{E}[\Lambda] = \frac{\alpha_0}{\beta_0},$$

the posterior mean has the linear form

$$\mathbb{E}[\Lambda \,|\, \{N_{1:k} = n_{1:k}\}] = c\frac{\alpha_0}{\beta_0} + (1-c)\frac{1}{k}\sum_{i=1}^{k} n_i,$$

where $c = \frac{\beta_0}{k+\beta_0}$. The higher the value of β_0 the more the posterior distribution is influenced by the prior expectation. That is not such surprising if we make us aware of the variance of the Gamma distribution (Remark B.5), which is

$$\mathbb{V}ar[\Lambda] = \frac{\alpha_0}{\beta_0^2}.$$

A high value of β_0 implies a high prior precision about possible values of Λ, thus a large sample size is needed to influence that belief.
The update process of the posterior parameters for $k = 1, 5, 10$ observations is illustrated in Figure 4.11 A. The corresponding update of the predictive distribution is in Figure 4.11 B and is calculated according Lemma 4.3.5, which says: The predictive distribution of a Poisson distribution on basis of k observations is a negative binomial distribution with α_k successes and success probability $\beta_k/(\beta_k + 1)$.

Lemma 4.3.5. *Let $\Lambda \sim \mathcal{G}amma(\alpha_0, \beta_0)$ be the prior distribution and $N_1, \ldots, N_k$ a sequence of i.i.d. random variable with $N_1 \,|\, \{\Lambda = \lambda\} \sim Pois(\lambda)$. Then the predictive distribution $N_{k+1} \,|\, \{N_{1:k} = n_{1:k}\}$ is a negative binomial distribution, i.e.,*

$$N_{k+1} \,|\, \{N_{1:k} = n_{1:k}\} \sim \mathcal{NB}\left(\alpha_k, \frac{\beta_k}{\beta_k + 1}\right),$$

where α_k and β_k are defined as in Lemma 4.3.4.

Proof. Let $\pi_\lambda(\cdot)$ denote the prior distribution of Λ. To determine the predictive distribution we use the law of total probability, which yields, cf. Lemma 4.3.4,

$$\begin{aligned}
\mathbb{P}(N_{k+1} = n_{k+1} \,|\, \{N_{1:k} = n_{1:k}\}) &= \int_0^\infty \mathbb{P}(N_{k+1} = n_{k+1} \,|\, \lambda)\pi_\lambda(\lambda \,|\, \{N_{1:k} = n_{1:k}\})d\lambda \\
&= \int_0^\infty \frac{\lambda^{n_{k+1}}}{n_{k+1}!}e^{-\lambda}\frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)}\lambda^{\alpha_k - 1}e^{-\beta_k\lambda}d\lambda \\
&= \frac{\beta_k^{\alpha_k}}{n_{k+1}!\,\Gamma(\alpha_k)}\int_0^\infty \lambda^{\alpha_k + n_{k+1} - 1}e^{-(\beta_k+1)\lambda}d\lambda \\
&= \frac{\beta_k^{\alpha_k}}{n_{k+1}!\,\Gamma(\alpha_k)}\frac{\Gamma(\alpha_k + n_{k+1})}{(\beta_k + 1)^{\alpha_k + n_{k+1}}} \\
&= \frac{\Gamma(\alpha_k + n_{k+1})}{n_{k+1}!\,\Gamma(\alpha_k)}\left(\frac{\beta_k}{\beta_k + 1}\right)^{\alpha_k}\left(\frac{1}{\beta_k + 1}\right)^{n_{k+1}}, \qquad (4.6)
\end{aligned}$$

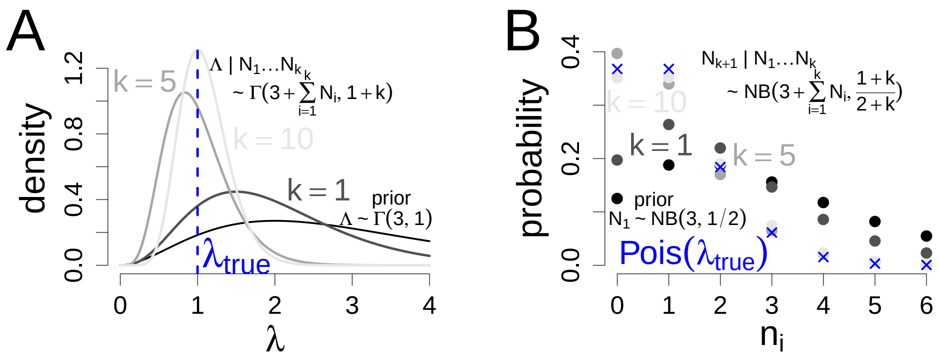


Figure 4.11: A. Update process of the posterior parameters α_k and β_k of the rate variable Λ for $k = 1, 5, 10$ observations. The true (unknown) rate parameter is $\lambda_{true} = 1$. As prior parameters we choose $\alpha_0 = 3$ and $\beta_0 = 1$. Thus starting with a broad $\mathcal{G}amma(3, 1)$-prior-distribution, we get closer to the true rate parameter λ_{true} by updating the posterior parameters α_k and β_k according to Lemma 4.3.4. B. The predictive distribution of $N_{k+1} \,|\, \{N_{1:k}\}$ for the same parameters and observations as in A. Thus starting with $\mathcal{NB}(3, 1/2)$ prior predictive distribution the posterior predictive distribution converges to a $Pois(1)$-distribution for a increasing number of observations k.

which is the probability mass function of a negative-binomial distribution with parameters α_k and $\beta_k/(\beta_k + 1)$, cf. Definition B.9. □

Remark 1: If we start with α_0 integer, the normalization factor in Equation (4.6) can be replaced by the binomial coefficient $\binom{\alpha_k + n_{k+1} - 1}{n_{k+1}}$, which is a more common representation of the negative binomial distribution with the interpretation: We encounter exactly n_{k+1} failures, before we encounter α_k successes, see also Remark B.10.

Remark 2: For a parametric model we know in general that the Bayesian updating yields asymptotically the true parameter, if the prior distribution does not exclude the true value. Here we can easily verify, that the predictive distribution converges to a $Pois(\lambda)$ distribution: With the law of large numbers we know

$$\frac{\alpha_k}{\beta_k} \xrightarrow{k \to \infty} \lambda$$

4.3.2.2 Application of BOCD

With the knowledge about the predictive distribution we are able to apply the BOCD to detect changes in the rate parameter.
First we consider a no change point scenario and compare the robustness of the rate for various expected number of spikes per oscillation cycle. Thereby we vary the prior parameter α_0 ($\beta_0 = 1$) of the gamma distribution to compare a low, mid and high prior belief about the rate. Second we consider those prior beliefs in an one change point scenario and analyze the change detection for a jump in the rate from 1 to 2 or from 1 to 4.
In each scenario we start with an uniform change point prior ($a_0 = b_0 = 1$).

No change point We consider various number of cycles $K = 10, 20, \dots, 500$ containing no change point and and various fixed rate parameters $\lambda = 1, 2, 4$, i.e., in each oscillation cycle k we observe

$$N_k \sim Pois(\lambda), \quad \text{and } N_k, N_{k'} \text{ independent } \forall\; k \neq k'.$$

In Figure 4.12 the number of detected change points is shown dependent on the number of observed cycles K. In Figure 4.12 A we choose a Gamma$(2, 1)$ prior distribution, which

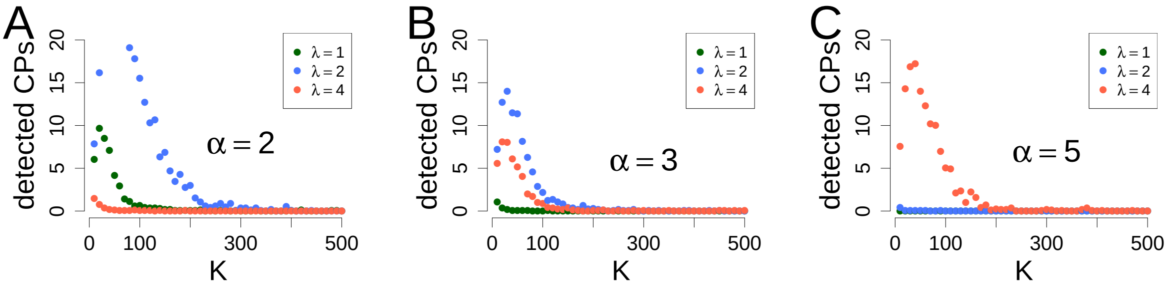


Figure 4.12: No change point scenario: We consider $K = 10, 20, \dots, 500$ independent Poisson(λ) distributed numbers and apply the BOCD with an uniform change point prior. We analyze three different rates $\lambda = 1$ (green), $\lambda = 2$ (blau) and $\lambda = 4$ (red) and three different prior beliefs. A. Small rate prior: $\alpha_0 = 2$ and $\beta_0 = 1$. B. Middle rate prior: $\alpha_0 = 3$ and $\beta_0 = 1$. C. High rate prior: $\alpha_0 = 5$ and $\beta_0 = 1$.

represents a small rate prior belief (expectation equals 2); in Figure 4.12 B we choose a Gamma$(3, 1)$ prior distribution, which represents a middle rate prior belief (expectation equals 3); in Figure 4.12 C we choose a Gamma$(5, 1)$ prior distribution, which represents a high rate prior belief (expectation equals 5).
Interestingly the small rate prior works well for a high rate (red), but bad for a small rate of 2 (blue), which equals the expectation. Equally, a high rate prior works well for small rates (green and blue), but bad for a high rate of 5 (red), which equals the expectation. In both cases we need at least $K = 200$ oscillation cycles to have a robust detection for all rates. A middle rate prior works adequate for all rates and with $K = 100$ oscillation cycles we have almost no false detections.
For example, a high rate prior detects many false change points in case of a high rate (at least in the short run), as it has high and not flat prior density in the true rate range and since we start with an uniform change point prior, there is some chance for many small changes, c.f. Section 4.3.3.2.

One change point We consider various numbers $K = 10, 20, \dots, 100$ of cycles containing exactly one change point in the mid of the sequence, i.e.,

$$N_1, \dots, N_{K/2} \sim Pois(1) \quad \text{and} \quad N_{K/2+1}, \dots, N_K \sim Pois(\lambda), \quad \text{and } N_1, \dots, N_K \text{ independent},$$

where we analyze the two cases $\lambda = 2$ (green) and $\lambda = 4$ (blue). In Figure 4.13 the number of all detected change points (dots) and the number of correctly detected change points (triangles) are shown dependent on the number of observed oscillation cycles K. Again in Figure 4.13 A

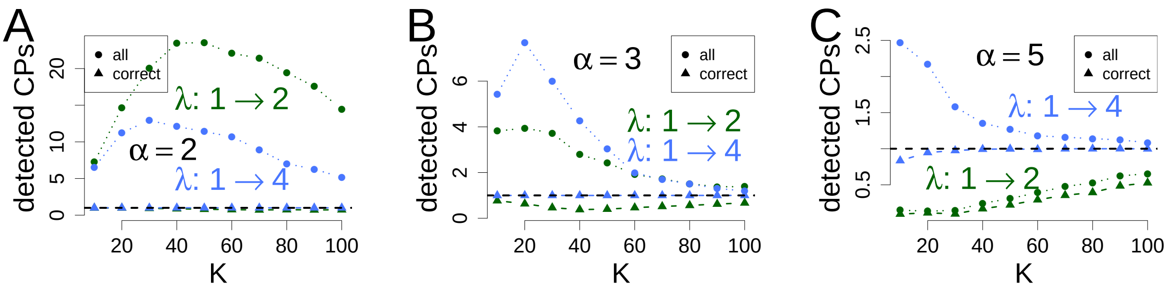


Figure 4.13: One change point scenario: We observe $K = 10, 20, \ldots, 100$ cycles, where in the first $K/2$ cycles we draw independent Poisson(1) distributed numbers and in the second $K/2$ cycles we draw independent Poisson(λ) distributed numbers, where we consider the two cases $\lambda = 2$ (green) and $\lambda = 4$ (blue). We apply the BOCD with an uniform change point prior and distinguish the number of all detected change points (dots) and only the number of correct detected change points (triangles). A. Small rate prior: $\alpha_0 = 2$ and $\beta_0 = 1$. B. Middle rate prior: $\alpha_0 = 3$ and $\beta_0 = 1$. C. High rate prior: $\alpha_0 = 5$ and $\beta_0 = 1$.

we choose a Gamma$(2, 1)$ prior distribution, in Figure 4.13 B we choose a Gamma$(3, 1)$ prior distribution and in Figure 4.13 C we choose a Gamma$(5, 1)$ prior distribution. Using a small rate prior we almost ever detect the true change point, but we make many false detections, even for $K = 100$ oscillation cycles (cf. Figure 4.13 A). A mid rate prior detects almost all changes from 1 to 4 (blue) and also robust for $K = 100$. Changes from 1 to 2 (green) can be detected appropriately for at least $K = 60$ and robust for $K = 100$. With a high rate prior a change detection operates very well for a change from 1 to 4 (blue) and is robust also for a change from 1 to 2 (green), but detects less true change points than a mid rate prior.
In summary with respect to the results of the no and one change point scenario and as the relevant rate parameters in the data set are in the range of 1 to 4, $\alpha_0 = 3$ seems an appropriate choice for the prior distribution of the rate, see also Section 4.3.3.2.

4.3.3 Rate and phase analysis

In this section we consider change points in the rate and phase simultaneously and question if the bivariate analysis can improve the change point detection compared to a pure rate analysis. In general, the change point detection should improve as only common change points in rate and phase are assumed (Zimek et al., 2012; Alippi et al., 2016). Nevertheless, if an additional change point in the phase can clearly improve the change detection, is an interesting question, as the changes are so small relative to its precision.
To apply the BOCD we need to determine a conjugate prior distribution on rate and phase. That can be done straightforward by combining the single prior distribution of rate and phase as we assumed independence, see Section 4.3.3.1.
Furthermore, to apply the BOCD we have to choose prior parameters for the change point probability and the rate and phase. In Section 4.3.3.2 we connect to the results of Section 4.3.1.2 and Section 4.3.2.2 and justify the prior choice of rate and phase for the ongoing studies. For a preferably broad view, we choose a uniform prior distribution on the change point probability (i.e. $a_0 = b_0 = 1$).

Afterwards we compare a bivariate rate and phase analysis with a pure rate analysis and a pure phase analysis. Therefore, in Section 4.3.3.3 we consider an one change point scenario and apply the BOCD and the BOCD with online decision (Section 4.2.4).

4.3.3.1 Conjugate prior and predictive distribution

For technical reasons we choose the same prior distribution for rate and phase as in a pure rate or phase code and assume independence of Λ and Φ, i.e.,

$$\pi(\lambda, \varphi) = \pi_{\alpha_0,\beta_0}(\lambda)\, \pi_{\mu_0,\tau_0}(\varphi),$$

where $\pi_{\alpha_0,\beta_0}(\cdot)$ is the density of a $\mathcal{G}amma(\alpha_0, \beta_0)$-distribution and $\pi_{\mu_0,\tau_0}(\cdot)$ the density of a $\mathcal{N}(\mu_0, \tau_0^2)$-distribution. With that choice the choice of the prior parameters for rate and phase is the same in the bivariate and in the univariate case.

As according to our sampling model, the spike times and spike numbers are drawn independently, i.e.,

$$p(n, \bar{x} \,|\, \Phi = \varphi, \Lambda = \lambda) = \begin{cases} \frac{\lambda^n}{n!} e^{-\lambda} \frac{1}{\sqrt{2\pi/n}} e^{-\frac{(\bar{x}-\varphi)^2}{2/n}}, & \text{if } n > 0, \\ \frac{\lambda^n}{n!} e^{-\lambda}, & \text{if } n = 0. \end{cases}$$

also the posterior distributions of rate and phase are conditionally independent, i.e.,

$$\pi(\lambda, \varphi \,|\, N = n, \bar{X} = \bar{x}) = \pi_{\alpha_0,\beta_0}(\lambda \,|\, N = n)\, \pi_{\mu_0,\tau_0}(\varphi \,|\, N = n, \bar{X} = \bar{x}).$$

Hence in the bivariate case we get the same posterior distributions for each component and obviously the same predictive distributions. With Equation (4.4) we can update the prior distribution of the phase, and with Equation (4.5) we can update the prior distribution of the rate, which yields also in the bivariate case an efficient application of the BOCD by updating conjugate prior distributions.

4.3.3.2 Prior parameter choice and behavior without change points

Here we motivate our choice of prior parameters of the BOCD in the change point model (Section 4.1) using the conjugate prior distributions from Section 4.3.3.1. We draw on the insights about prior parameters for the phase (Section 4.3.1.2 and 4.3.1.4) and the rate (Section 4.3.2.2) and compare our choices. Specifically, we illustrate that the choice of a relatively uninformative prior can reduce the probability of falsely estimating change points, which is crucial for the performance of the algorithm. For the change point probability η itself we choose a non-informative uniform prior, i.e., $a_0 = b_0 = 1$.

The parameters for the conjugate prior distributions are chosen such that the prior distributions carry relatively little information, i.e., that the parameters are approximately uniformly distributed in the range found in the experimental data (cmp. Section A, $\varphi \in [0, 0.75]$, $\lambda \leq 4$), while having less mass outside this range. For the mean and variance of the phase parameter we choose $\mu_0 = 0$ and $\tau_0^2 = 4$ (Figure 4.14 B, green). The phase parameter is set to $\varphi = 0$ in the simulations as the results are shift invariant and the prior density hardly depends on φ for $\varphi \in [0, 0.75]$. Figure 4.14 A-E shows that a more informative prior with smaller variance $\tau_0^2 = 1$ (blue) can sometimes massively overestimate the change point probability, while this is not the case for the prior with higher variance (panel E). For the rate prior, we choose $\alpha_0 = 3$

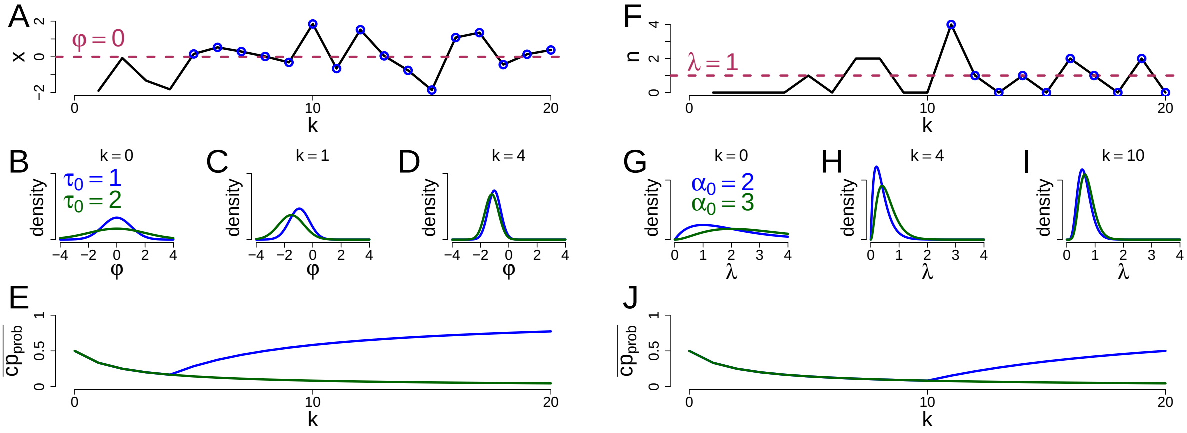


Figure 4.14: Illustration of performance of BOCD in the setting of no change points, for two different prior distributions of the phase (A-E, prior of the phase is normal $\mathcal{N}(0, \tau_0^2)$) and the rate (F-J, prior of the rate is $\mathcal{Gamma}(\alpha_0, 1)$). First row (A,F): Example sequence of observed spike times ($X_i \sim \mathcal{N}(0,1)$, A) or observed numbers of spikes ($N_i \sim Pois(1)$, B) and the detected change points (circles) as a function of the observed cycle k. The blue prior distributions (with $\tau_0 = 1$ on the left and $\alpha_0 = 2$ on the right) falsely detect change points (blue circles), while the green prior distributions (with $\tau_0 = 2$ on the left and $\alpha_0 = 2$ on the right) do not falsely detect change points. The red dotted line indicates the true phase (A) or rate (F) parameter. Second line indicates prior ($k = 0$) and posterior distributions for phase (B-D) and rate (G-I) parameters. Third row: The resulting predictive distribution that a change point occurs for the different priors, indicated by colors. For the blue prior, the change point probability is increased abruptly after the first detection and grows steadily.

and $\beta_0 = 1$ (Figure 4.14 F-J, green). Again, a more informative prior, e.g., with parameter $\alpha_0 = 2$ (blue) can sometimes lead to a strong overestimation of the change point probability.

The examples shown in Figure 4.14 are supported in systematic simulations shown in Figure 4.15 in a setting without change points. For the phase parameter, the smaller variance $\tau_0^2 = 1$ (panel A, circles) yielded a strongly increased number of falsely detected change points, while this was not the case for the less informative prior $\tau_0^2 = 4$ (triangles). We therefore consider the case $\tau_0^2 = 4$ for the bivariate analysis. For the rate parameter, the choice of $\alpha_0 = 2$ yielded an increased number of falsely detected change points for small and medium rates (panel B, circles) as compared to $\alpha_0 = 3$ (triangles). In case of high rates ($\lambda = 4$) the choice $\alpha_0 = 2$ can increase the number of detected change points, and we therefore consider $\alpha_0 \in \{2, 3\}$ in the bivariate analysis, in which rate and phase parameters are considered for change point detection (panel C). For $\alpha_0 = 3$ (triangles) we observe almost no falsely detected change points even for a small number of cycles. Even for $\alpha_0 = 2$ (circles), the number of falsely detected change points is only slightly increased. This suggests that the bivariate change point detection is highly robust against random deviations within the time series, which may in the univariate case evoke falsely detected change points.

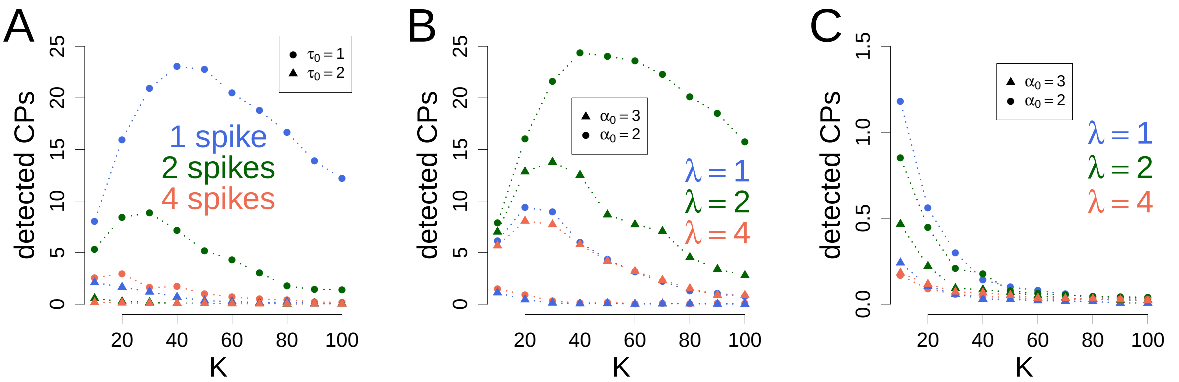


Figure 4.15: Number of detected change points as a function of the total number K of oscillation cycles (1000 simulations per data point). A. Pure phase code for 1, 2 or 4 spikes per oscillation cycle. Dots represent a $\mathcal{N}(0,1)$ prior distribution, triangles a $\mathcal{N}(0,4)$ prior distribution for the phase parameter φ. B. Pure rate code for the rates $\lambda \in \{1,2,4\}$ indicated by colors and for the prior distribution $\mathcal{G}amma(2,1)$ (dots) and $\mathcal{G}amma(3,1)$ (triangles). C. Bivariate analysis with rate and phase parameters and the prior parameters $\tau_0^2 = 4$ and $\alpha_0 = 3$ (dots) and $\alpha_0 = 2$ (triangles).

4.3.3.3 Behavior with one change in rate and phase

The simulation results of Figure 4.14 and 4.15 suggest that the use of bivariate information can decrease the number of falsely detected change points. Now we investigate the behavior in a setting with exactly one change point. Specifically, we compare the univariate and bivariate analysis both in the BOCD and in its extended version with online decision. As both the rate and the phase parameter carry information about the change point, we expect an improved change point detection in the bivariate analysis (Zimek et al., 2012; Alippi et al., 2016). However, as changes in the phase parameter are relatively small as compared to its precision, we are specifically interested in quantifying the amount of improvement by the phase parameter in comparison to a pure rate analysis. Therefore, we consider the scenario of one change point of a fixed magnitude occurring in both parameters exactly in the middle of the time horizon. We choose parameter settings corresponding to plausible neuronal parameter ranges, letting the rate increase from 1 to 2 or from 1 to 4 and the phase increase from 0 to 0.5.

BOCD

Figures 4.16 A and B show two examples of BOCD analysis applied to spike trains consisting of $K = 50$ cycles with a phase change from 0 to 0.5 and a rate change from 1 to 2 (A) and 1 to 4 (B) at time $K/2$. The detected change points of the BOCD analysis are shown as red dashed vertical lines, using only the rate parameter (first row), only the phase parameter (second row) and the bivariate analysis (third row). In panel A, the BOCD based on either rate or phase alone could not locate the true change point. The rate-BOCD even falsely detected a change point at the end of the sequence. In panel B, again the phase-BOCD did not detect any change point, and the rate-BOCD detected four change points of which only one is close to

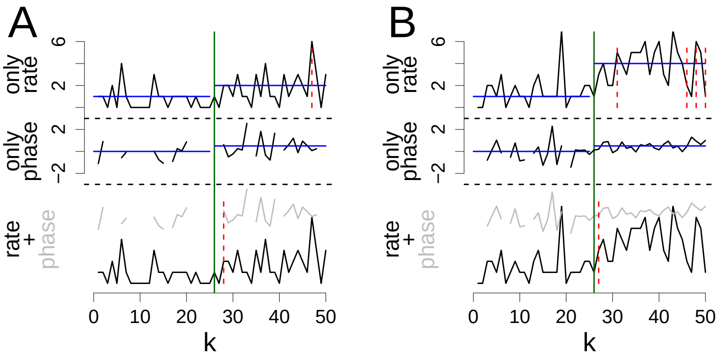


Figure 4.16: Example sequences of observed cycles, with a change point in the middle (green). The phase changes from 0 to 0.5, the rate changes from 1 to 2 (A) or from 1 to 4 (B). We apply the BOCD with univariate analyses using only rate (first row, black), only phase (second row, mean spike time shown in black) and the bivariate analysis (third row, spike time shown in gray, spike numbers in black). True rate or phase is shown in blue, estimated change points as red dashed lines.

the true change point. In both examples, the bivariate analysis estimated exactly one change point, which was very close to the true change point.

These results are supported in systematic simulations shown in Figure 4.17, where we investigate the number of correctly and of falsely detected change points for different time horizons $K = 10$ to $K = 100$ and the scenario of exactly one change point at $K/2$. A change point is called

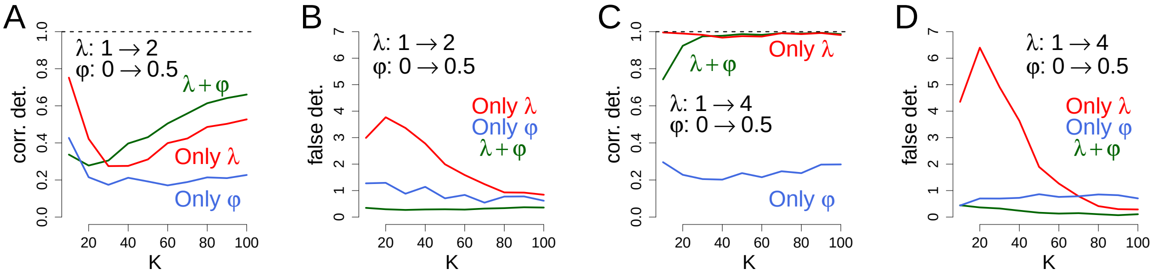


Figure 4.17: Evaluation of BOCD using bivariate analysis in the one change point setting with a change point in the middle. The phase changes from 0 to 0.5, the rate changes from 1 to 2 (A,B) or from 1 to 4 (C,D). Average fraction of correctly detected change points (A,C) and average number of falsely detected change points (B,D) for sequences of length $K = 10$ to $K = 100$, 1000 simulations per data point. Pure rate analysis is shown in red, pure phase analysis in blue, and bivariate analysis in green.

correctly detected here if its distance to the estimated change point is less than 3. Otherwise it is called falsely detected. Similar to the examples in Figure 4.16, Figure 4.17 shows that the rate-BOCD (red) shows high detection rates for the true change points, but is however likely to strongly overestimate the number of falsely detected change points, specifically for small time horizons. While the BOCD based on phase alone (blue) shows only a low percentage of correctly detected change points and also increased numbers of falsely detected change points, the BOCD based on rate and phase parameters (green) shows both high sensitivity to true change points and robustness against random fluctuations in the time series.

BOCD with online decision

In order to apply the BOCD with online decision, cf. Section 4.2.4, we consider spike trains of length $K = 100$ with exactly one change point in the middle and consider different decision delays d (Figure 4.18). Concerning correctly detected change points (Figure 4.18 A and C),

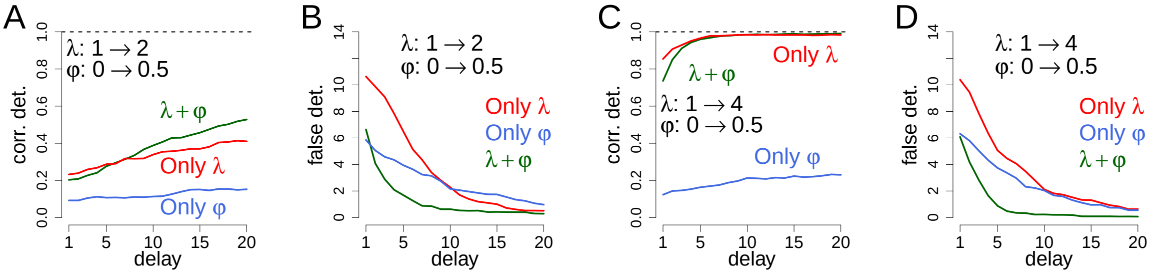


Figure 4.18: Evaluation of BOCD with online decision using bivariate analysis in the one change point setting with a change point in the middle. The phase changes from 0 to 0.5, the rate changes from 1 to 2 (A,B) or from 1 to 4 (C,D). Average fraction of correctly detected change points (A,C) and average number of falsely detected change points (C,D) for sequences with length $K = 100$, 1000 simulations per data point. Pure rate analysis is shown in red, pure phase analysis in blue, and bivariate analysis in green.

we find that the detection probability with online decision is considerably smaller than when using the whole spike train, and the number of correctly detected change points increases with the delay. Concerning falsely detected change points (Figure 4.18 B and D), a univariate analysis using rate shows the largest number of false alarms. The bivariate analysis shows a higher amount of correctly detected change points and a smaller amount of falsely detected change points than the univariate analysis, particularly for small delays.

These results suggest that imprecise phases can increase the number of correctly detected change points, while considerably decreasing the number of falsely detected change points as well as decreasing the delay required for correct change point detection and thus, increasing the speed and robustness of change point detection.

4.3.3.4 $\bar{X}$ is not appropriate for change point detection

In Section 4.3.1.3 to calculate the predictive distribution, we update the posterior distribution of Φ successively in the cycle, i.e.,

$$p\left(X_{1:n_{k+1}}^{(k+1)} = x_{1:n_{k+1}}^{(k+1)} \,|\, \bar{X}_{1:k} = \bar{x}_{1:k}\right) = p\left(X_1^{(k+1)} \,|\, \bar{X}_{1:k} = \bar{x}_{1:k}\right)$$
$$\cdot \ldots \cdot p\left(X_{n_{k+1}}^{(k+1)} \,|\, X_{1:n_{k+1}-1}^{(k+1)} = x_{1:n_{k+1}-1}^{(k+1)}, \bar{X}_{1:k} = \bar{x}_{1:k}\right).$$

First we motivate, why this is necessary, and we can not only consider the mean spike time in each cycle. Second we analyze, how not successive updating or only using the mean spike time affects the change point detection. For better readability we consider the first and second cycle starting with prior parameters μ_0, τ_0^2 and assume one spike in the first cycle (x_1) and two spikes $x_2 = (z_1, z_2)$ in the second cycle.
Note that the successive update only concerns the predictive distribution and not the posterior distribution, since the mean spike time is sufficient for φ, i.e.,

$$\Phi \,|\, \{Z_1 = z_1, \dots, Z_n = z_n\} \overset{d}{=} \Phi \,|\, \{\bar{Z} = \bar{z}, N = n\},$$

where $\bar{Z} := 1/n \sum_{i=1}^{n} Z_i$. Furthermore, note that the order of the successive update does not affect the predictive distribution $p(z_1, z_2 \,|\, X_1 = x_1)$, i.e.,

$$p(z_2 \,|\, Z_1 = z_1, X_1 = x_1) \cdot p(z_1 \,|\, X_1 = x_1) = p(z_1 \,|\, Z_2 = z_2, X_1 = x_1) \cdot p(z_2 \,|\, X_1 = x_1).$$

Let us first recall the BOCD and its basic concept of determining the distribution of the run length. According to Equation (4.1) the BOCD (for simplification with a known change point probability η) calculates at cycle $k = 2$ (note here we start with cycle $k = 1$ instead of $k = 0$)

$$\mathbb{P}(R_2 = 1 \,|\, X_{1:2} = x_{1:2}) = \frac{\overbrace{\mathbb{P}(R_2 = 1 \,|\, R_1 = 0)}^{\eta}\, p(x_2 \,|\, X_1 = x_1) p(x_1) \overbrace{\mathbb{P}(R_1 = 0)}^{1}}{p(x_{1:2})}$$

and

$$\mathbb{P}(R_2 = 0 \,|\, X_{1:2} = x_{1:2}) = \frac{\overbrace{\mathbb{P}(R_2 = 0 \,|\, R_1 = 0)}^{1-\eta}\, p(x_2 \,|\, X_1 = x_1) p(x_0) \overbrace{\mathbb{P}(R_1 = 0)}^{1}}{p(x_{1:2})}.$$

The quotient of both run lengths yields

$$\frac{\mathbb{P}(R_2 = 1 \,|\, X_{1:2} = x_{1:2})}{\mathbb{P}(R_2 = 0 \,|\, X_{1:2} = x_{1:2})} = \frac{\eta}{1-\eta} \frac{p(x_2 \,|\, X_1 = x_1)}{p(x_2)}. \tag{4.7}$$

Thus if the quotient with the predictive distributions in Equation (4.7) is different for the raw spike times $x_2 = (z_1, z_2)$ and the mean spike time $\bar{x}_2 := 1/2(z_1 + z_2)$, the BOCD does not calculates the same distribution of the run lengths in both cases and thus detects not the same change points in general.
So now we want to explore that in general

$$\frac{p(x_2 \,|\, X_1 = x_1)}{p(x_2)} \neq \frac{p(\bar{x}_2 \,|\, X_1 = x_1)}{p(\bar{x}_2)}. \tag{4.8}$$

Let us first consider the raw spike times. With Section 4.3.1.3 we write

$$Z_1 \,|\, \{X_1 = x_1\} \sim \mathcal{N}(\mu_1, \overbrace{\tau_1^2 + \sigma^2}^{\sigma_1^2}) \quad \text{and} \quad Z_2 \,|\, \{Z_1 = z_1, X_1 = x_1\} \sim \mathcal{N}(\mu_{11}, \overbrace{\tau_{11}^2 + \sigma^2}^{\sigma_{11}^2})$$

and

$$Z_1 \sim \mathcal{N}(\mu_0, \overbrace{\tau_0^2 + \sigma^2}^{\sigma_0^2}) \quad \text{and} \quad Z_2 \,|\, \{Z_1 = z_1\} \sim \mathcal{N}(\mu_{01}, \overbrace{\tau_{01}^2 + \sigma^2}^{\sigma_{01}^2}).$$

With that we write the left quotient of Equation (4.8) as

$$\begin{aligned}
&\frac{p(z_1, z_2 \,|\, X_1 = x_1)}{p(z_1, z_2)} \\
&= \frac{p(z_2 \,|\, Z_1 = z_1, X_0 = x_0) p(z_1 \,|\, X_1 = x_1)}{p(z_2 \,|\, Z_1 = z_1) p(z_1)} = \frac{\frac{1}{2\pi\sigma_{11}\sigma_1} \exp\left(-\frac{1}{2\sigma_{11}^2}(z_2 - \mu_{11})^2 - \frac{1}{2\sigma_1^2}(z_1 - \mu_1)^2\right)}{\frac{1}{2\pi\sigma_{01}\sigma_0} \exp\left(-\frac{1}{2\sigma_{01}^2}(z_2 - \mu_{01})^2 - \frac{1}{2\sigma_0^2}(z_1 - \mu_0)^2\right)} \\
&= \frac{\sigma_{01}\sigma_0}{\sigma_{11}\sigma_1} \exp\left(-\frac{1}{2\sigma_{11}^2}(z_2 - \mu_{11})^2 + \frac{1}{2\sigma_{01}^2}(z_2 - \mu_{01})^2 - \frac{1}{2\sigma_1^2}(z_1 - \mu_1)^2 + \frac{1}{2\sigma_0^2}(z_1 - \mu_0)^2\right).
\end{aligned}$$

Let us now consider the mean spike time ($n = 2$). Here we obtain with the above notations

$$\bar{X}_2 \,|\, \{X_1 = x_1\} \sim \mathcal{N}(\mu_1, \sigma_1^2/n) \quad \text{and} \quad \bar{X}_2 \sim \mathcal{N}(\mu_0, \sigma_0^2/n)$$

and for the right quotient of Equation (4.8)

$$\begin{aligned}
\frac{p(\bar{x}_2 \,|\, X_1 = x_1)}{p(\bar{x}_2)} &= \frac{\frac{1}{\sqrt{2\pi\sigma_1^2/n}} \exp\left(-\frac{n}{2\sigma_1^2}(\bar{x}_2 - \mu_1)^2\right)}{\frac{1}{\sqrt{2\pi\sigma_0^2/n}} \exp\left(-\frac{n}{2\sigma_0^2}(\bar{x}_2 - \mu_0)^2\right)} \\
&= \frac{\sigma_0}{\sigma_1} \exp\left(-\frac{n}{2\sigma_1^2}(\bar{x}_2 - \mu_1)^2 + \frac{n}{2\sigma_0^2}(\bar{x}_2 - \mu_0)^2\right),
\end{aligned}$$

where we can already see due to the different quotients in front of the exponential function that the case of raw spikes and the mean spike time have different quotients of predictive distributions and thus different run length distributions. For sure, we give a short example. Let $x_1 = 1$ and $z_1 = 0$ and $z_2 = 5.4$. Furthermore, let $\mu_0 = 0$ and $\tau_0^2 = 4$ (and $\sigma^2 = 1$). This yields

$$\begin{aligned}
\mu_1 &= \frac{1}{1/4 + 1} = \frac{4}{5} &&\text{and} & \tau_1^2 &= \frac{1}{1/4 + 1} = \frac{4}{5} \\
\mu_{11} &= \frac{1}{1/4 + 2} = \frac{4}{9} &&\text{and} & \tau_{11}^2 &= \frac{1}{1/4 + 2} = \frac{4}{9} \\
\mu_{01} &= 0 &&\text{and} & \tau_{01}^2 &= \frac{1}{1/4 + 1} = \frac{4}{5}
\end{aligned}$$

and overall

$$\frac{p(x_2 \,|\, X_1 = x_1)}{p(x_2)} \approx 1.043 \neq 0.964 \approx \frac{p(\bar{x}_2 \,|\, X_1 = x_1)}{p(\bar{x}_2)}.$$

With a change point probability $\eta = 1/2$ we would decide at cycle $k = 2$ for a change point in case of the mean spike time, but not in case of the raw spike times. The mean spike time is more vulnerable to unlikely events than the raw spike times, as the difference in the quotients increases for unlikely events, i.e., for $z_2 = 10$ (all other equal) we obtain

$$\frac{p(x_2 \,|\, X_1 = x_1)}{p(x_2)} \approx 0.034 \neq 0.014 \approx \frac{p(\bar{x}_2 \,|\, X_1 = x_1)}{p(\bar{x}_2)},$$

which is a difference of factor 2.43. Figure 4.19 confirms our expectation that the mean spike time will detect falsely (and correctly) more change points. We consider $K = 10, 20, \ldots, 100$

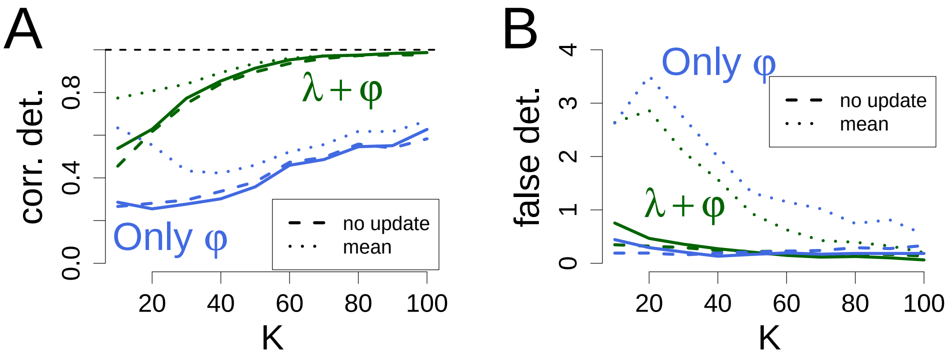


Figure 4.19: Comparison of BOCD using raw spike times (line) to the BOCD using the mean spike times (dashed lines) in the one change point setting with a change point in the middle. The phase changes from 0 to 0.5, the rate changes from 2 to 4. Average fraction of correctly detected change points (A) and average number of falsely detected change points (B) for sequences of length $K = 10$ to $K = 100$, 1000 simulations per data point. Pure phase analysis is shown in blue, and bivariate analysis in green. The dashed lines represent the BOCD using raw spike times but without successive update in each cycle.

cycles with one change point at $K/2$, where the rate jumps from 2 to 4 and the phase from 0 to 0.5. We choose our usual prior parameters of $\alpha_0 = 3$, $\beta_0 = 1$, $\mu_0 = 0$, $\tau_0^2 = 4$ and $a_0 = b_0 = 1$. If we apply the BOCD to the mean spike time (dotted line), the number of correctly detected change points is higher (Figure 4.19 A), but only due to the excessive higher number of falsely detected change points (Figure 4.19 B). This holds for a pure phase analysis (blue) and the bivariate analysis (green). Interestingly, if we do not successively update the parameters in one cycle and falsely calculate $p(z_1, \ldots, z_n \,|\, X_1 = x_1) = p(z_1 \,|\, X_1 = x_1) \cdots p(z_n \,|\, X_1 = x_1)$, we obtain almost the same number of correctly and falsely detected change points (dashed lines).

4.4 Mixture of conjugate prior distributions

In this section we discuss an approach to consider special information about the stimulus specific rate and phase parameters. Yet we assumed a wide prior distribution, both for rate and phase, and search for any changes in rate or phase without any dependence up to a simultaneous appearance. But what if we have special knowledge about the stimuli structure? For example, we know stimuli are either coded by small rates or by high rates, but not by middle rates. Or we know that small rates are only connected with small phases and high rates with high phases.

We can not implement such information with a wide conjugate prior distribution. Nevertheless, due to computational cost we need an easy posterior update process, comparable with conjugate distributions. A solution is a mixture distribution of conjugate prior distributions (Diaconis and Ylvisaker, 1985). The clue is that the posterior distribution of a mixture of conjugate distributions is again a mixture of conjugate distributions, see Section 4.4.1. So we can draw on the results of the conjugate distributions of rate and phase and use the same formulas to update the posterior and predictive distributions, cf. Section 4.3.1 and 4.3.2. Additionally, we need to determine the posterior weights of the mixture, which can be done efficiently. Furthermore, in Section 4.4.1 we present the coin toss example, for which it is plausible that any prior information can be represented with conjugate prior distributions. In general this result holds for exponential family sample distributions, see Proposition 4.4.6.

To apply the BOCD we notice in Section 4.4.2 that the predictive distribution based on the mixture distribution is just a mixture of the single weighted predictive distributions. Furthermore, we specify how to adjust the BOCD in context of a mixture distribution.

In Section 4.4.3 we apply the setting of a mixture distribution to rate and phase and $S = 2$ types of stimuli. Due to the specialized information for the ranges of the rate or phase parameter, we need to adjust the change point prior parameters in the BOCD, otherwise an overestimation of change points in the short run occurs. However also in case of specific information about stimuli properties the phase contains additional information compared to a pure rate analysis and improves the change point detection.

4.4.1 Update process mixture distributions

In this section we take up the concept of mixtures of conjugate distribution (Diaconis and Ylvisaker, 1985). To motivate the use of mixture distributions in general, we start with a coin toss example. Afterwards we define a mixture distribution in general (Definition 4.4.2). With Lemma 4.4.3 we have a guidance how to calculate the posterior mixture distribution efficiently, if we start with a mixture of conjugate prior distributions. Subsequently we apply that to the coin toss example and calculate the posterior mixture distribution. In the coin toss example it is plausible, that if we mixture many specialized prior distributions, any distribution on $(0, 1)$ can be approximated with any precision. In general this holds for any exponential family sample distribution, see Proposition 4.4.6.

Example 4.4.1. *Let us consider a coin toss Y with k throws and random probability of success H, i.e., given $\{H = \eta\}$ is*

$$Y \sim Bin(k, \eta).$$

Assume we know that the coin is biased, but we do not know in which direction. So we are sure that either H is small or H is large. We have the option to throw the coin k times. Our task is now, how to choose the prior distribution to get an appropriate posterior belief about H? With Claim 4.2.1 we already know that $H \sim Beta(a_0, b_0)$ is a conjugate prior distribution. If we believe H is rather small, we would maybe choose $H \sim Beta(10, 30)$ (see Figure 4.20 A, blue dashed line), if we believe H is rather large, we would maybe choose $H \sim Beta(30, 10)$ (see Figure 4.20 A, green dashed line). As both are equal likely for us, we just mix both distributions, each with weight $1/2$ (see Figure 4.20 A, black line), i.e.,

$$\pi_\eta(\cdot) = \frac{1}{2}\beta(\cdot, 10, 30) + \frac{1}{2}\beta(\cdot, 30, 10),$$

where $\beta(\cdot, a_0, b_0)$ is the density of a beta-distribution with parameter (a_0, b_0).

The posterior update of a mixture of conjugate prior distributions of Example 4.4.1 continues in Example 4.4.5.

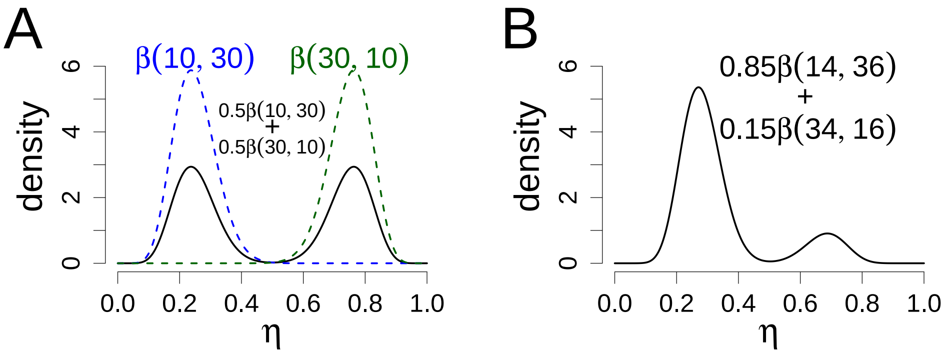


Figure 4.20: Coin toss example. A. The blue dashed line represents the prior belief of a small success probability, the green dashed line the prior belief of a high success probability. If both possibilities are equally likely, we mix both distributions with probability 1/2. The result is the black line. B. The updated posterior mixture distribution after observing only 4 successes in 10 throws.

Definition 4.4.2. *A mixture distribution $\pi(\cdot)$ is any convex combination of other distributions $\pi_s(\cdot)$, i.e.,*

$$\pi(\cdot) := \sum_{s=1}^{S} \omega_s \pi_s(\cdot), \quad \textit{with} \ \sum_{s=1}^{S} \omega_s = 1, \, \omega_s \geq 0, \, S > 1.$$

The following Lemma states that the posterior distribution of a mixture of conjugate prior distributions is again a mixture of conjugate distributions.

Lemma 4.4.3. *Consider a realization $X = x$ of a Bayesian Model $\mathcal{B}(\Pi, \mathcal{P})$ with $\pi^{(0)}(\cdot) := \pi(\cdot) = \sum_{s=1}^{S} \omega_s^{(0)} \pi_s^{(0)}(\cdot)$ a mixture of conjugate prior distributions $\pi_s^{(0)}(\cdot)$, $s = 1, \ldots, S$. Furthermore let $\pi_s^{(1)}(\cdot) := \pi_s^{(0)}(\cdot \,|\, X = x)$ be the single posterior distributions of a Bayesian Model $\mathcal{B}(\Pi_s, \mathcal{P})$. Then the posterior mixture distribution is given by*

$$\pi^{(1)}(\theta) := \pi^{(0)}(\theta \,|\, X = x) = \sum_{s=1}^{S} \omega_s^{(1)} \pi_s^{(1)}(\theta),$$

where

$$\omega_s^{(1)} := \frac{\omega_s^{(0)} c_s}{\sum_{s=1}^{S} \omega_s^{(0)} c_s} \quad \textit{with} \quad c_s := \int_{-\infty}^{\infty} p_{\tilde{\theta}}(x) \pi_s^{(0)}(\tilde{\theta}) \, d\tilde{\theta}.$$

Proof. Let $p_\theta(\cdot)$ be the sampling distribution and $\pi_s^{(0)}(\cdot)$, $s = 1, \ldots, S$, conjugate prior distributions. Consider the update process using Bayes' rule: Given an observation x, the posterior distribution $\pi_s^{(1)}(\cdot)$ is given by

$$\pi_s^{(1)}(\theta) := \pi_s^{(0)}(\theta \,|\, x) = \frac{\pi_s^{(0)}(\theta) p_\theta(x)}{c_s}, \qquad s = 1, \ldots, S,$$

with

$$c_s := \int_{-\infty}^{\infty} p_{\tilde{\theta}}(x)\pi_s^{(0)}(\tilde{\theta})\, d\tilde{\theta}.$$

Now let us consider a mixture of conjugate distributions

$$\pi^0(\theta) = \sum_{s=1}^{S} \omega_s^{(0)} \pi_s^{(0)}(\theta).$$

The posterior distribution $\pi^{(1)}(\cdot)$ of this mixture distribution is

$$\begin{aligned}
\pi^{(1)}(\theta) := \pi^{(0)}(\theta \,|\, x) &= \frac{\pi^{(0)}(\theta) p_\theta(x)}{\int_{-\infty}^{\infty} p_{\tilde{\theta}}(x)\pi^{(0)}(\tilde{\theta})\, d\tilde{\theta}} \\
&= \frac{\sum_{s=1}^{S} \omega_s^{(0)} \pi_s^{(0)}(\theta) p_\theta(x)}{\int_{-\infty}^{\infty} p_{\tilde{\theta}}(x) \sum_{s=1}^{S} \omega_s^{(0)} \pi_s^{(0)}(\theta)\, d\tilde{\theta}} \\
&= \frac{\sum_{s=1}^{S} \omega_s^{(0)} c_s \pi_s^{(1)}(\theta)}{\sum_{s=1}^{S} \omega_s^{(0)} c_s} \\
&= \sum_{s=1}^{S} \omega_s^{(1)} \pi_s^{(1)}(\theta),
\end{aligned}$$

where

$$\omega_s^{(1)} = \frac{\omega_s^{(0)} c_s}{\sum_{s=1}^{S} \omega_s^{(0)} c_s} \quad \text{with} \quad \sum_{s=1}^{S} \omega_s^{(1)} = 1.$$

Hence the posterior distribution is again a mixture of conjugate distribution, only the weights changed. □

Remark 4.4.4. *As we know the explicit form of the posterior distributions $\pi_s^{(1)}(\cdot)$ we can determine the constants c_s without integrating by reverting Bayes' rule, i.e.,*

$$c_s = \frac{\pi_s^{(0)}(\theta)}{\pi_s^{(1)}(\theta)} p_\theta(x),$$

for a θ with $\pi_s^{(0)}(\theta) > 0$, $\pi_s^{(1)}(\theta) > 0$ and $p_\theta(x) > 0$.

With Lemma 4.4.3 and Remark 4.4.4 we are able to calculate the posterior distribution of a mixture of conjugate distributions computationally efficient and nevertheless are very flexible to represent our prior information. For clarification, we continue with Example 4.4.1.

Example 4.4.5. *Let us follow the setting of Example 4.4.1, so we choose the conjugate prior distributions*

$$\pi_1^{(0)}(\cdot) = \beta(\cdot, 10, 30) \quad \textit{and} \quad \pi_2^{(0)}(\cdot) = \beta(\cdot, 30, 10),$$

where $\beta(\cdot, a_0, b_0)$ is the density of a beta-distribution with parameter (a_0, b_0) and the weights

$$\omega_1^{(0)} = \frac{1}{2} \quad \textit{and} \quad \omega_2^{(0)} = \frac{1}{2}.$$

Assume we throw the coin 10 times and observe only 4 times success. Thus the single posterior distributions are

$$\pi_1^{(1)}(\cdot) = \beta(\cdot, 10+4, 30+6) \quad \textit{and} \quad \pi_2^{(1)}(\cdot) = \beta(\cdot, 30+4, 10+6)$$

and with $\eta = 1/2$ the constants can be calculated as (Remark 4.4.4)

$$c_1 = \frac{\pi_1^{(0)}(\eta)}{\pi_1^{(1)}(\eta)}\, \mathbb{P}(Y=4) \approx 0.14 \quad \textit{and} \quad c_1 = \frac{\pi_2^{(0)}(\eta)}{\pi_2^{(1)}(\eta)}\, \mathbb{P}(Y=4) \approx 0.02,$$

where $Y \sim Binom(10, \eta)$. Finally the new weights are

$$\omega_1^{(1)} = \frac{\omega_1^{(0)} c_1}{\omega_1^{(0)} c_1 + \omega_2^{(0)} c_2} \approx 0.85 \quad \textit{and} \quad \omega_2^{(1)} = \frac{\omega_2^{(0)} c_2}{\omega_1^{(0)} c_1 + \omega_2^{(0)} c_2} \approx 0.15.$$

Thus the posterior mixture distribution is

$$\pi^{(1)} = 0.85 \cdot \beta(\cdot, 10+4, 30+6) + 0.15 \cdot \beta(\cdot, 30+4, 10+6),$$

which reflect our strong concern that the success probability is rather small. An graphical illustration of the posterior distribution can be found in Figure 4.20 B.

The mixture of conjugate prior distributions has especially its charm, if we think of S stimuli types and one randomly chosen stimulus. For each stimulus type s we have some prior information about the distribution of Θ_s and we roughly know the probability of each stimulus type (so we can set the weights $\omega_s^{(0)}$). By mixture, we can construct an appropriate prior distribution, which moreover can be calculated efficiently.

In general mixture of conjugate prior distributions are useful, if we have very specialized prior information, which can not be represented by a conjugate prior distribution. In the light of the beta distribution it is plausible, that any distribution on $(0, 1)$ can be approximated by mixture of beta distributions: If we choose the parameters a and b of a beta distribution large enough, the density is close to a point mass at $a/(a + b)$ and with many beta distributions, we can approximate any distribution on $(0, 1)$.

More general in case of an exponential family distribution any prior distribution can be well approximated by a finite mixture of conjugate prior distributions, cf. Proposition 4.4.6.

Proposition 4.4.6. *Let Ω_θ be the natural parameter space of a d-dimensional exponential family distribution. For any probability distribution $\pi(\cdot)$ on Ω_θ and any $\epsilon > 0$ there are weights ω_s and (k_s, t_s), $k_s > 0$ and t_s lies in the interior of the convex hull of the support of the measure μ, and $m < \infty$ such that if*

$$\tilde{\pi}(\theta) = \sum_{s=1}^{S} \omega_s c(k_s, t_s) exp\left(\theta^T (k_s t_s) - k_s A(\theta)\right)$$

then

$$d(\pi, \tilde{\pi}) < \epsilon,$$

where d is the Lévy-Prokhorov metric.

Proof. See (Diaconis and Ylvisaker, 1985). □

Remark 4.4.7. *In a separable metric space convergence of measures in the Lévy-Prokhorov metric is equivalent to weak convergences of measures.*

Remark 4.4.8. *Again the intuition of Proposition 4.4.6 is that we choose k_s very large to have the single conjugate priors concentrate at its mean and the mixture of point masses converge to the prior distribution $\pi(\cdot)$.*

4.4.2 Predictive distribution and adjustment of BOCD

To apply the BOCD we also need the predictive distribution: Based on each conjugate prior distribution itself we know the predictive distribution. To determine the predictive distribution of the mixture, we just need to weight the single predictive distributions by the posterior weights ω_s, cf. Lemma 4.4.9. Afterwards we specify, how to adjust the BOCD in context of a mixture distribution.

Lemma 4.4.9. *Consider a Bayesian Model $\mathcal{B}(\Pi, \mathcal{P})$ with a mixture of conjugate prior distributions $\pi^{(0)}(\cdot) := \pi(\cdot) = \sum_{s=1}^{S} \omega_s^{(0)} \pi_s^{(0)}(\cdot)$. Furthermore, let given $\{\Theta = \theta\}$ be $X, X' \sim P_\theta$ and independent and let $\rho_s(\cdot)$, $s = 1, \dots, S$, denote the single predictive distribution of the Bayesian Model $\mathcal{B}(\Pi_s, \mathcal{P})$.*
Then the predictive distribution $\rho(\cdot) := p(\cdot \mid X' = x')$ is given by

$$\rho(x) = \sum_{s=1}^{S} \omega_s^{(1)} \rho_s(x),$$

where

$$\omega_s^{(1)} := \frac{\omega_s^{(0)} c_s}{\sum_{s=1}^{S} \omega_s^{(0)} c_s} \quad \text{with} \quad c_s := \int_{-\infty}^{\infty} p_{\tilde{\theta}}(x) \pi_s^{(0)}(\tilde{\theta})\, d\tilde{\theta}.$$

Proof. As

$$\rho(x) = \int_{-\infty}^{\infty} p_\theta(x) \pi^{(0)}(\theta \mid \{X' = x'\})\, d\theta$$

and from Lemma 4.4.3 we know that $\pi^{(0)}(\theta \mid \{X' = x'\}) = \sum_{s=1}^{S} \omega_s^{(1)} \pi_s^{(0)}(\theta \mid \{X' = x'\})$, thus

$$\begin{aligned}
\rho(x) &= \int_{-\infty}^{\infty} p_\theta(x) \sum_{s=1}^{S} \omega_s^{(1)} \pi_s^{(0)}(\theta \mid \{X' = x'\})\, d\theta \\
&= \sum_{s=1}^{S} \omega_s^{(1)} \int_{-\infty}^{\infty} p_\theta(x) \pi_s^{(0)}(\theta \mid \{X' = x'\})\, d\theta \\
&= \sum_{s=1}^{S} \omega_s^{(1)} \rho_s(x).
\end{aligned}$$

□

Algorithm. Basically to adjust the BOCD to the setting of a mixture distribution, the determination of the predictive distribution is more complicated: We need to introduce weights ω_s, which need to be updated for each possible path. Thus step 1, 3 and 7 of the algorithm need to be adjusted. Let $\pi_1(\cdot), \ldots, \pi_S(\cdot)$ denote the conjugate prior distributions and $p_1(\cdot), \ldots, p_S(\cdot)$ the predictive distributions.

1. Initialize (choose prior parameters of Θ and change point probability H)

$$\mathbb{P}(R_0 = 0) := 1;\ a_0 := a_0;\ b_0 := b_0$$

$$n_0^{(0)} := \left(n_0^{(0,1)}, \ldots, n_0^{(0,S)}\right);\ t_0^{(0)} := \left(t_0^{(0,1)}, \ldots, t_0^{(0,S)}\right);\ \omega_0^{(0)} := \left(\omega_0^{(0,1)}, \ldots, \omega_0^{(0,S)}\right).$$

3. Evaluate Predictive Probability (for $j = 0, \ldots, k$)

$$\psi_k^{(j)} := \omega_k^{(j,1)} p_1\left(x_k \,|\, n_k^{(j,1)}, t_k^{(j,1)}\right) + \cdots + \omega_k^{(j,S)} p_S\left(x_k \,|\, n_k^{(j,S)}, t_k^{(j,S)}\right).$$

7. Update sufficient statistics (for $j = 1, \ldots, k$ and $s = 1, \ldots, S$)

$$\begin{aligned} n_{k+1}^{(0,s)} &= n_0^{(0,i)} && \text{and} \quad n_{k+1}^{(j,s)} = n_k^{(j-1,s)} + 1 \\ t_{k+1}^{(0,s)} &= n_0^{(0,s)} t_0^{(0,s)} && \text{and} \quad t_{k+1}^{(j,s)} = t_k^{(j-1,s)} + t(x_k) \\ \omega_{k+1}^{(0,s)} &= \omega_0^{(0,s)} && \text{and} \quad \omega_{k+1}^{(j,s)} = \frac{c_k^{(j-1,s)} \omega_k^{(j-1,s)}}{\sum_{\tilde{s}=1}^{S} c_k^{(j-1,\tilde{s})} \omega_k^{(j-1,\tilde{s})}} \end{aligned}$$

where

$$c_k^{(j-1,s)} = \int_{-\infty}^{\infty} p_\theta(x_k) \pi_s\left(\theta \,|\, n_k^{(j-1,s)}, t_k^{(j-1,s)}\right) d\theta.$$

4.4.3 Application BOCD - special rate and phase information

First we discuss how to choose a mixture of prior distributions for rate and phase, if we have specific prior information. Afterwards we apply the BOCD with various prior parameters. Assume we have specific prior information, and we know, there are only small rates or high rates and equally there are only small phases or high phases. Additional we know there are only two cases, how stimuli are coded:

1. small rate $\longleftrightarrow$ small phase
2. high rate $\longleftrightarrow$ high phase.

To create a corresponding mixture of conjugate prior distributions, we draw on the results of Section 4.3.1 and 4.3.2:
For $N_i \sim Pois(\Lambda)$, $i = 1, \ldots k$,

$$\text{if} \quad \Lambda \sim \mathcal{G}amma(\alpha_0, \beta_0), \quad \text{then} \quad \Lambda \,|\, \{N_{1:k} = n_{1:k}\} \sim \mathcal{G}amma\left(\alpha_0 + \sum_{i=1}^{k} n_i, \beta_0 + k\right)$$

and for $X_i \sim \mathcal{N}(\Phi, 1)$, $i = 1, \ldots k$,

$$\text{if} \quad \Phi \sim \mathcal{N}\left(\mu_0, \tau_0^2\right), \quad \text{then} \quad \Phi \,|\, \{X_{1:k} = x_{1:k}\} \sim \mathcal{N}\left(\mu_k, \tau_k^2\right),$$

where

$$\varphi_k := \frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum_{i=1}^k x_i}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{k}{\sigma^2}} \quad \text{and} \quad \tau_k^2 := \frac{1}{\frac{1}{\tau_0^2} + \frac{k}{\sigma^2}}.$$

Let $\phi_{\mu_0,\tau_0}(\cdot)$ denote the density of a normal distribution with mean μ_0 and standard deviation τ_0 and $g_{\alpha_0,\beta_0}(\cdot)$ the density of a gamma distribution with parameter α_0 and β_0. To consider the prior information of only small or high rates or phases, we could choose a mixture of two conjugate priors, i.e.,

$$\pi_\varphi^{(0)}(\cdot) = \frac{1}{2}\phi_{\mu_{01},\tau_{01}}(\cdot) + \frac{1}{2}\phi_{\mu_{02},\tau_{02}}(\cdot)$$

and

$$\pi_\lambda^{(0)}(\cdot) = \frac{1}{2}g_{\alpha_{01},\beta_{01}}(\cdot) + \frac{1}{2}g_{\alpha_{02},\beta_{02}}(\cdot)$$

For $\mu_{01} = 0$, $\tau_{01} = 0.1$ and $\varphi_{02} = 0.5$, $\tau_{02} = 0.1$ the update process of phase mixture is illustrated in Figure 4.21 A, the update process of rate mixture for $\alpha_{01} = 8$, $\beta_{01} = 4$ ($\mathbb{E}[\Lambda] = 2$) and $\alpha_{02} = 32$, $\beta_{02} = 8$ ($\mathbb{E}[\Lambda] = 4$) in Figure 4.21 B.

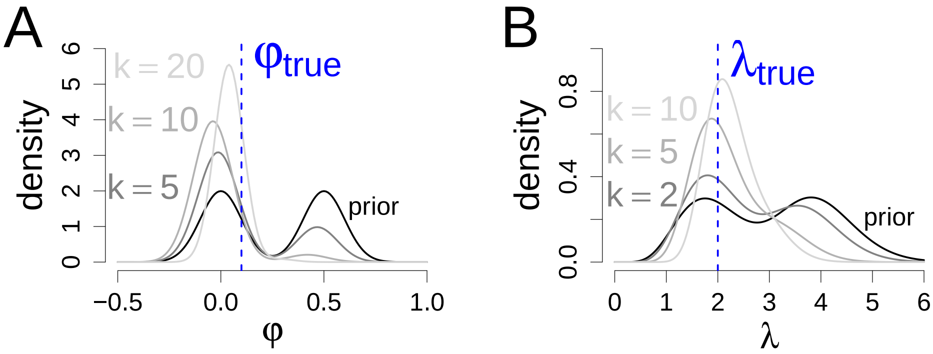


Figure 4.21: A. Update process of a phase mixture. As prior distribution we choose a mixture (weights equal 0.5) of two normal distributions with mean 0 and 0.5 and both with standard deviation 0.1. The posterior distributions after observing $k = 5, 10, 20$ realizations shifts more and more to the true phase of 0.1. B. Update process of a rate mixture. As prior distribution we choose a mixture (weights equal 0.5) of two gamma distributions with shape 8 and 32 and rate 4 and 8. The posterior distributions after observing $k = 2, 5, 10$ realizations shifts more and more to the true rate of 2.

To consider the prior information of the connection between rate and phase (small with small, high with high), we choose the following joint mixture distribution

$$\pi_{\lambda,\varphi}^{(0)}(\cdot) = \frac{1}{2}\phi_{\mu_{01},\tau_{01}}(\cdot)g_{\alpha_{01},\beta_{01}}(\cdot) + \frac{1}{2}\phi_{\mu_{02},\tau_{02}}(\cdot)g_{\alpha_{02},\beta_{02}}(\cdot).$$

Each component is a conjugate distribution (cf. Section 4.3.3.1), thus with Lemma 4.4.3 we can determine the posterior mixture distribution and with Lemma 4.4.9 the predictive

distribution. Thus we can apply the BOCD efficiently. Again we consider the following change point setting:
Let a change point in rate and phase occur exactly at the mid of the sequence $K/2$ (K even), i.e

$$N_k \sim Pois(2), \quad k = 1, \dots, K/2$$

and

$$N_k \sim Pois(4), \quad k = K/2 + 1, \dots, K$$

and for each $j = 1, \dots, N_k$

$$X_j^{(k)} \sim \mathcal{N}(0, 1), \quad k = 1, \dots, K/2$$

and

$$X_j^{(k)} \sim \mathcal{N}\left(0.75, 1\right), \quad k = K/2 + 1, \dots, K.$$

We consider two prior cases: Either we are sure that there are no mid rates or phases, we call this precise prior information, i.e.,

$$\begin{aligned} \alpha_0 = (64, 256) \quad &\text{and} \quad \beta_0 = (32, 64) \\ \mu_0 = (0, 0.75) \quad &\text{and} \quad \tau_0 = (0.1, 0.1), \end{aligned}$$

see Figure 4.22 A and B (purple line) or we have imprecise prior information

$$\begin{aligned} \alpha_0 = (16, 64) \quad &\text{and} \quad \beta_0 = (8, 16) \\ \mu_0 = (0, 0.75) \quad &\text{and} \quad \tau_0 = (0.2, 0.2), \end{aligned}$$

see Figure 4.22 A and B (orange line). In both cases the mixture prior information is to specialized compared to a wide prior choice (cf. Section 4.3.3.2) that an uniform change point prior ($a_0 := b_0 := 1$) results in a strong overestimation of the number of change points in the short run. So here we start with the expectation of one change point within the sequence ($a_0 := 1;\ b_0 := K$).
Again we compare the three variants of a pure rate, a pure phase or the bivariate analysis.

Results - Figure 4.23:
First we notice that the more precise prior information (Figure 4.23 C and D) only slightly differs from a less precise prior information (Figure 4.23 A and B). However the precise prior information results in a higher number of correctly detected change points and does less false detections, especially in the short run ($K < 50$).
Here in the setting of specialized prior information a pure rate (red) and a pure phase (blue) analysis basically show the same change detection ability: Both types of analysis have almost the same number of correctly detected change points and the same number of falsely detected change points. But that is no surprise if we look at both prior distributions (Figure 4.22). The change from rate 2 to 4 is separated in the rate prior the same as a change from 0 to 0.75 in the phase prior. So we should also see the same ability in the change point detection, if we observe such changes. So the question, if small changes in the phase can help to improve the

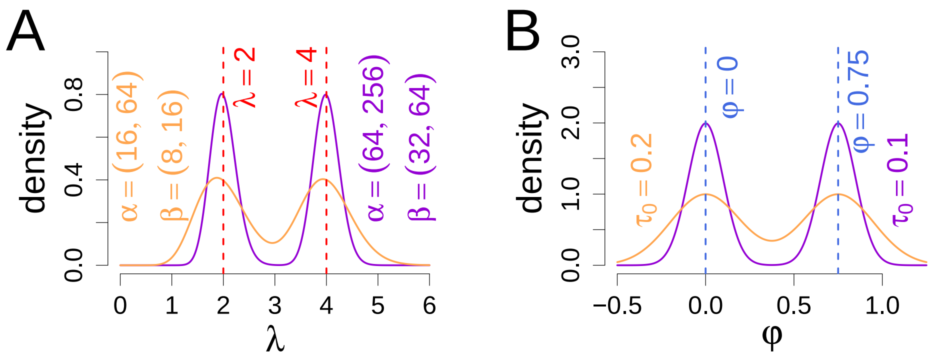


Figure 4.22: Precise and imprecise mixture prior information for rate and phase. A. Mixture of two gamma distributions, precise (purple line) $\alpha_0 = (64, 256)$, $\beta_0 = (32, 64)$, imprecise (orange line) $\alpha_0 = (16, 64)$, $\beta_0 = (8, 16)$. B. Mixture of two normal distributions, precise (purple line) $\mu_0 = (0, 0.75)$, $\tau_0 = (0.1, 0.1)$, imprecise (orange line) $\mu_0 = (0, 0.75)$, $\tau_0 = (0.2, 0.2)$

.

detection ability of the rate, depends in addition to the phase change crucial on the precision of our prior information.

However, using rate and phase (green line) simultaneously results again in an improved change point detection: We detect a higher number of change points correctly, and we do less false detections. Thus the phase can improve the change detection based on a pure rate analysis also in case of precise prior information.

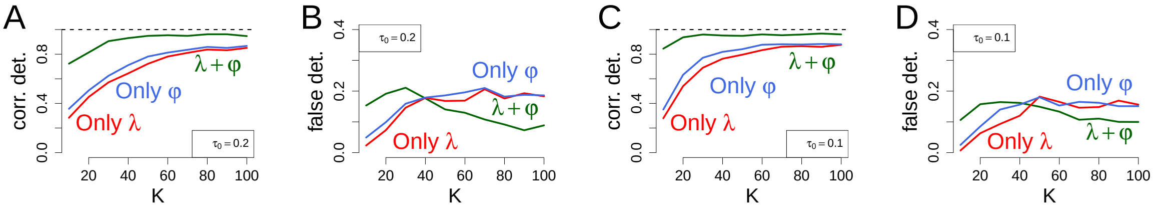


Figure 4.23: Evaluation of BOCD using prior mixture distributions in the one change point setting with a change point in the middle. The phase changes from 0 to 0.75, the rate changes from 2 to 4. Average fraction of correctly detected change points (A,C) and average number of falsely detected change points (B,D) for sequences of length $K = 10$ to $K = 100$, 1000 simulations per data point. A and B. Precise prior information of $\alpha_0 = (64, 256)$, $\beta_0 = (32, 64)$ and $\mu_0 = (0, 0.75)$, $\tau_0 = (0.1, 0.1)$. C and D. Imprecise prior information of $\alpha_0 = (16, 64)$, $\beta_0 = (8, 16)$ and $\mu_0 = (0, 0.75)$, $\tau_0 = (0.2, 0.2)$. Pure rate analysis is shown in red, pure phase analysis in blue, and bivariate analysis in green.

4.5 Unknown spike precision and changes in σ^2

Up to this point we assumed that we know the precision σ^2 of the spike times $X_j^{(k)}$, i.e., given $\{A_k = a_k\}$ change points up to time k we have $\Phi_{a_k} \sim \mathcal{N}(\mu_0, \tau_0^2)$ and given $\{\Phi_{a_k} = \varphi_{a_k}\}$ we choose

$$X_1^{(k)}, \ldots, X_{n_k}^{(k)} \sim \mathcal{N}(\varphi_{a_k}, \sigma^2) \text{ independently,}$$

where σ^2 is fixed and known. In short we write $X_1^{(k)}, \ldots, X_{n_k}^{(k)} \sim \mathcal{N}(\Phi_{A_k}, \sigma^2)$. As the distinction of two phase parameters φ_1 and φ_2 only depends on the quotients φ_1/σ and φ_2/σ, we chose w.l.o.g. $\sigma = 1$. But what if σ is unknown and can even change at some points?
So we assume in this section, we observe in one oscillation cycle

$$X_1^{(k)}, \ldots, X_{n_k}^{(k)} \sim \mathcal{N}(\Phi_{A_k}, \varsigma_{A_k}^2),$$

where Φ_{A_k} and $\varsigma_{A_k}^2$ are random variables. Or more formally we extend the change point model (Section 4.1) by an sequence $(\Phi_0, \varsigma_0^2), (\Phi_1, \varsigma_1^2), \ldots$ of phase and precision parameters with prior distribution $\pi_{\varphi,\sigma^2}(\cdot)$ and $(\varphi_0, \tilde{\sigma}_0^2), (\varphi_1, \tilde{\sigma}_1^2), \ldots$ a random realization. At every change point now a new realization of Λ, Φ and ς^2 is drawn. As a consequence in cycle k overall $N_k \sim Pois(\lambda_{A_k})$ spikes are chosen and are placed independently according to a $\mathcal{N}(\varphi_{A_k}, \tilde{\sigma}_{A_k}^2)$-distribution. So now the precision of the spike times is unknown and locked to the change point process. We use the joint notation of phase and precision $(\pi_{\varphi,\sigma^2}(\cdot))$ due to technical reasons, see Section 4.5.1 for a conjugate prior distribution.

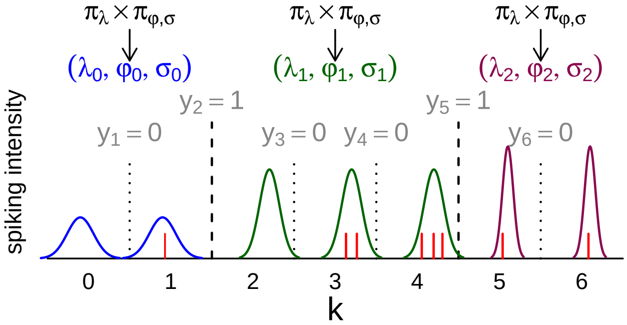


Figure 4.24: Change point model with random spike precision. We extend the model of Section 4.1: Now rate, phase and precision parameters can change in time as follows. For each oscillation cycle an independent Bernoulli random variable $Y_k \sim Ber(\eta)$ indicates whether a change point occurs. If no change point occurs ($Y_k = 0$), the rate λ, phase φ and precision σ^2 remain identical to the previous oscillation cycle ($k-1$). If a change point occurs ($Y_k = 1$), new parameters for rate λ, phase φ and precision σ^2 are chosen according to the prior distributions π_λ and $\pi_{\varphi,\sigma^2}(\cdot)$.

Our aim is to explore how an additional change in ς^2 impacts the change point detection on basis of a pure phase or a rate and phase analysis, see Section 4.5.5. Therefore we first research how changes in the spike precision affect the ability of the phase to detect change points, if we erroneously assume a constant spike precision σ^2. Second off we assume changes in rate, phase and spike precision occur simultaneously and consider the benefit of a simultaneous analysis of rate, phase and precision compared to a pure rate analysis.

Following the Bayesian procedure we need to specify a prior distribution $\pi_{\varphi,\sigma^2}(\cdot)$ on (Φ, ς^2). To apply the BOCD to the setting of an unknown spike precision with a numeric practical implementation, we need to determine a conjugate prior distribution on (Φ, ς^2). The conjugate prior distribution of a Normal distribution with random expectation and variance parameter is well known, i.e., see Gelman et al. (2013), but we show how the general results of Section 3.4.3 can be used to obtain the conjugate prior with a schematic workaround. Beneficial of this procedure is that we can transfer the general interpretation of the prior parameters k_0 (number of prior observations) and t_0 (prior belief of the sufficient statistic) to the specific prior distribution. Therefore, we notice in Section 4.5.1 that the sampling distribution is a 2-dimensional exponential family distribution and use the general result of Lemma 3.4.13 to construct a conjugate prior distribution by mimicking the likelihood.
In Section 4.5.2 we show how to set the prior parameters if we have prior expectation μ_0 for the phase parameter and σ_0^2 for the spike precision.
For a descriptive understanding of the posterior update process we determine in Section 4.5.3 the marginal distributions of Φ and ς^2.
To apply the BOCD to the setting of unknown phase and unknown precision, we determine the predictive distribution of $X_{k+1} \,|\, \{X_{1:k} = x_{1:k}\}$ in Section 4.5.4.

4.5.1 The exponential family approach

In Claim 4.5.1 we verify that the normal distribution is a 2-dimensional exponential family distribution. With that we can apply Lemma 3.4.13 to obtain a conjugate prior distribution by mimicking the likelihood, cf. Lemma 4.5.2. Equally, with Lemma 3.4.13 we are able to determine the posterior distribution, see Lemma 4.5.3.

Claim 4.5.1. *The normal distibution is a 2-dimensional exponential family distribution, i.e., for $X \sim \mathcal{N}(\varphi, \sigma)$ the density can be written as*

$$\phi_{\varphi,\sigma^2}(x) = \underbrace{\frac{1}{\sqrt{\pi}}}_{h(x)} \exp\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}^T \underbrace{\begin{pmatrix} x \\ x^2 \end{pmatrix}}_{t(x)} + \underbrace{\frac{1}{2}\log|\theta_2| + \frac{\theta_1^2}{4\theta_2}}_{-A(\theta)} \right),$$

where $\theta_1 := \varphi/\sigma^2$ and $\theta_2 := -(2\sigma^2)^{-1}$.

Proof. According to Definition 3.4.1 we verify, if the density of the normal distribution can be written as

$$\phi_{\varphi,\sigma^2}(x) = h(x) \cdot \exp(\theta^T t(x) - A(\theta)).$$

So we rewrite

$$\begin{aligned}\phi_{\varphi,\sigma^2}(x) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\varphi)^2}{2\sigma^2}\right)\\ &= \frac{1}{\sqrt{\pi}}\sqrt{(2\sigma^2)^{-1}}\exp\left(-(2\sigma))^{-1}x^2+\frac{\varphi}{\sigma^2}x-\frac{\varphi^2}{2\sigma^2}\right)\\ &= \frac{1}{\sqrt{\pi}}|\theta_2|^{1/2}\exp\left(\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}^T\begin{pmatrix}x\\ x^2\end{pmatrix}+\frac{\theta_1^2}{4\theta_2}\right)\\ &= \frac{1}{\sqrt{\pi}}\exp\left(\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}^T\begin{pmatrix}x\\ x^2\end{pmatrix}+\frac{1}{2}\log|\theta_2|+\frac{\theta_1^2}{4\theta_2}\right),\end{aligned}$$

where $\theta_1 := \varphi/\sigma^2$ and $\theta_2 := -(2\sigma^2)^{-1}$. □

Lemma 4.5.2. *Consider a Bayesian model with* $X \mid \{\Phi = \varphi, \varsigma^2 = \sigma^2\} \sim \mathcal{N}(\varphi, \sigma^2)$. *We obtain a conjugate prior distribution, if we choose*

$$\varsigma^2 \sim \mathcal{IG}(1/2(k_0+3), k_0(t_{02}-t_{01}^2)/2)$$

and given $\{\varsigma^2 = \sigma^2\}$

$$\Phi \sim \mathcal{N}(t_{01}, \sigma^2/k_0),$$

with $k_0 > 0$ *and* $t_{02} - t_{01}^2 > 0$.

Proof. As we are in the setting of a 2-dimensional exponential family model (Claim 4.5.1), according to Lemma 3.4.13 we obtain a conjugate prior distribution $\pi(\theta)$ if we mimic the likelihood, i.e.,

$$\pi(\theta) \sim \exp(\theta^T k_0 t_0 - k_0 A(\theta)),$$

with $A(\theta) := \frac{1}{2}\log|\theta_2| + \frac{\theta_1^2}{4\theta_2}$. To parametrise in terms of (φ, σ^2) we use the changes of variables formula, where

$$(\theta_1, \theta_2) = g^{-1}(\varphi, \sigma^2) := \left(\frac{\varphi}{\sigma^2}, (-2\sigma^2)\right).$$

The Jacobi matrix is given by

$$\mathcal{J}g^{-1}(\varphi, \sigma^2) = \begin{pmatrix}1/\sigma^2 & -\varphi/\sigma^4\\ 0 & (2\sigma^2)^{-1}\end{pmatrix}$$

and the determinant of the Jacobi matrix is

$$det\ \mathcal{J}g^{-1}(\varphi, \sigma^2) = \frac{1}{4\sigma^6}.$$

Overall the change of variables formula yields

$$\begin{aligned}
\pi(\varphi,\sigma^2) &= det\ \mathcal{J}g^{-1}(\varphi,\sigma^2)\cdot\pi\left(g^{-1}(\varphi,\sigma^2)\right)\\
&\sim \frac{1}{4\sigma^6}\exp\left(k_0\begin{pmatrix}\varphi/\sigma^2\\(-2\sigma^2)^{-1}\end{pmatrix}^T\begin{pmatrix}t_{01}\\t_{02}\end{pmatrix}+\frac{k_0}{2}\log\left((2\sigma^2)^{-1}\right)+k_0\frac{\varphi^2/\sigma^4}{4(-2\sigma^2)^{-1}}\right)\\
&\sim \frac{1}{\sigma^6}(\sigma^2)^{-k_0/2}\exp\left(k_0\frac{\varphi}{\sigma^2}t_{01}-k_0\frac{t_{02}}{2\sigma^2}-\frac{k_0}{2}\frac{\varphi^2}{\sigma^2}\right)\\
&\sim \left(\sigma^2\right)^{-1/2}\exp\left(-\frac{k_0\varphi^2-2k_0\varphi t_{01}+k_0t_{01}^2}{2\sigma^2}\right)\left(\sigma^2\right)^{-(k_0+5)/2}\exp\left(\frac{k_0t_{01}^2-k_0t_{02}}{2\sigma^2}\right)\\
&\sim \underbrace{\left(\sigma^2\right)^{-1/2}\exp\left(-\frac{k_0}{2\sigma^2}(\varphi-t_{01})^2\right)}_{\sim\,\mathcal{N}(t_{01},\sigma^2/k_0)}\underbrace{\left(\sigma^2\right)^{-(k_0+5)/2}\exp\left(-\frac{k_0(t_{02}-t_{01}^2)}{2\sigma^2}\right)}_{\sim\,\mathcal{IG}((k_0+3)/2,k_0(t_{02}-t_{01}^2)/2},
\end{aligned}$$

which is proportional to the product of a $\mathcal{N}(t_{01},\sigma^2/k_0)$-distribution and an $\mathcal{IG}((k_0{+}3)/2,k_0(t_{02}-t_{01}^2)/2)$-distribution, cf. Definition B.11. □

Lemma 4.5.3. *Consider a Bayesian model with* $X_j\,|\,\{\Phi=\varphi,\varsigma^2=\sigma^2\}\sim\mathcal{N}(\varphi,\sigma^2)\ \forall\, j=1,\ldots,k$ *and* $X_1,\ldots,X_k$ *conditional independent. Furthermore, let*

$$\varsigma^2\sim\mathcal{IG}(1/2(k_0+3),k_0(t_{02}-t_{01}^2)/2)$$

and given $\{\varsigma^2=\sigma^2\}$

$$\Phi\sim\mathcal{N}(t_{01},\sigma^2/k_0).$$

Then the posterior distributions are

$$\varsigma^2\,|\,\{X_{1:k}=x_{1:k}\}\sim\mathcal{IG}\left(\frac{k_0+k+3}{2},\frac{1}{2}\left(k_0t_{02}+k\overline{x^2}-\frac{1}{k_0+k}\left(k_0t_{01}+k\bar{x}\right)^2\right)\right),$$

and

$$\Phi\,|\,\{X_{1:k}=x_{1:k},\varsigma^2=\sigma^2\}\sim\mathcal{N}\left(\frac{k_0}{k_0+k}t_{01}+\frac{k}{k_0+k}\bar{x},\frac{\sigma^2}{k_0+k}\right),$$

with $\bar{x}:=1/k\sum_{i=1}^k x_i$ *and* $\overline{x^2}:=1/k\sum_{i=1}^k x_i^2$.

Proof. According to Lemma 3.4.13 the posterior distribution is given by

$$\pi(\theta\,|\,X_{1:k}=x_{1:k})\sim\exp\left(\theta^T(k_0t_0+\sum_{i=1}^k t(x_i))-(k_0+k)A(\theta)\right),$$

where $t(x_i)=(x_i,x_i^2)^T$, $(\theta_1,\theta_2)=\left(\frac{\varphi}{\sigma^2},\left(-2\sigma^2\right)\right)$ and $A(\theta)=-1/2\log|\theta_2|-\theta_1^2/(4\theta_2)$.

Parametrization in terms of φ and σ yields

$$\begin{aligned}
&\pi(\varphi,\sigma \,|\, X_{1:k}=x_{1:k})\\
&= \mathit{det}\ \mathcal{J}g^{-1}(\varphi,\sigma^2)\cdot\pi\left(g^{-1}(\varphi,\sigma^2)\,|\,x_1,\dots,x_k\right)\\
&\sim \frac{1}{\sigma^6}\exp\left(\begin{pmatrix}\varphi/\sigma^2\\ (-2\sigma^2)^{-1}\end{pmatrix}^T\begin{pmatrix}k_0t_{01}+k\bar{x}\\ k_0t_{02}+k\bar{x^2}\end{pmatrix}+\frac{k_0+k}{2}\log\left((2\sigma^2)^{-1}\right)-\frac{k_0+k}{2}\frac{\varphi^2}{\sigma^2}\right)\\
&\sim \frac{1}{\sigma^6}(\sigma^2)^{-(k_0+k)/2}\exp\left(\frac{\varphi}{\sigma^2}(k_0t_{01}+k\bar{x})-\frac{k_0t_{02}+k\bar{x^2}}{2\sigma^2}-\frac{k_0+k}{2}\frac{\varphi^2}{\sigma^2}\right)\\
&\sim \left(\sigma^2\right)^{-1/2}\exp\left(-\frac{(k_0+k)\left(\varphi^2-2\varphi\left(\tfrac{k_0}{k_0+k}t_{01}+\tfrac{k}{k_0+k}\bar{x}\right)+\left(\tfrac{k_0}{k_0+k}t_{01}+\tfrac{k}{k_0+k}\bar{x}\right)^2\right)}{2\sigma^2}\right)\\
&\cdot\left(\sigma^2\right)^{-(k_0+k+5)/2}\exp\left(\frac{(k_0+k)\left(\tfrac{k_0}{k_0+k}t_{01}+\tfrac{k}{k_0+k}\bar{x}\right)^2-k_0t_{02}-k\bar{x^2}}{2\sigma^2}\right)\\
&\sim \underbrace{\left(\sigma^2\right)^{-1/2}\exp\left(-\frac{k_0+k}{2\sigma^2}\left(\varphi-\left(\frac{k_0}{k_0+k}t_{01}+\frac{k}{k_0+k}\bar{x}\right)\right)^2\right)}_{\sim\,\mathcal{N}(k_0/(k_0+k)t_{01}+k/(k_0+k)\bar{x},\sigma^2/(k_0+k))}\\
&\underbrace{\cdot\left(\sigma^2\right)^{-(k_0+k+5)/2}\exp\left(-\frac{(k_0+k)\left(\tfrac{k_0}{k_0+k}t_{02}+\tfrac{k}{k_0+k}\bar{x^2}-\left(\tfrac{k_0}{k_0+k}t_{01}+\tfrac{k}{k_0+k}\bar{x}\right)^2\right)}{2\sigma^2}\right)}_{\sim\,\mathcal{IG}((k_0+k+3)/2,1/2(k_0t_{02}+k\bar{x^2}-1/(k_0+k)(k_0t_{01}+k\bar{x})^2))},
\end{aligned}$$

cf. Definition B.11. □

4.5.2 Appropriate prior parameters

From Lemma 4.5.2 and 4.5.3 we know a conjugate prior and posterior distribution of (Φ,ς^2), but we still need a guidance to choose the prior parameters k_0 and t_0 appropriate. The variable k_0 can be interpreted as prior sample size and determines the precision of our prior belief. How to choose t_0, if we have prior expectation μ_0 for the phase parameter and σ_0^2 for the spike precision is stated in Lemma 4.5.4. How this choice affects the parameters of the posterior distribution is declared in Corollary 4.5.5. An illustration of the prior and posterior distribution of (Φ,ς^2) is shown in Figure 4.25.

Lemma 4.5.4. *Consider a Bayesian model with $X\,|\,\{\Phi=\varphi,\varsigma^2=\sigma^2\}\sim\mathcal{N}(\varphi,\sigma^2)$. We obtain a conjugate prior distribution with*

$$\mathbb{E}[\Phi]=\mu_0 \quad \textit{and} \quad \mathbb{E}[\varsigma^2]=\sigma_0^2$$

if we choose

$$\varsigma^2\sim\mathcal{IG}\left(\gamma_0,(\gamma_0-1)\sigma_0^2\right)$$

with $\gamma_0 := (k_0+3)/2$ *and given* $\{\varsigma^2 = \sigma^2\}$

$$\Phi \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{k_0}\right),$$

with $k_0 > 0$.

Proof. According to Fisher-Neyman factorization theorem and Claim 4.5.1 $t(x) = (x, x^2)$ is a sufficient statistic for $\theta = (\theta_1, \theta_2)$ and due to the simple variable transformation it is also sufficient for φ and σ^2. So regarding Lemma 4.5.2 and as

$$\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}\left[X \mid \Phi\right]\right] = \mathbb{E}[\Phi] = \mu_0,$$

we choose

$$t_{01} = \mathbb{E}[X] = \mu_0.$$

Furthermore we note

$$\mathbb{E}\left[\mathbb{V}ar\left[X \mid \Phi, \varsigma^2\right]\right] = \mathbb{E}\left[\varsigma^2\right] = \sigma_0^2$$

and using the variance decomposition formula yields

$$\begin{aligned} t_{02} = \mathbb{E}\left[X^2\right] &= \mathbb{V}ar[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}\left[\mathbb{V}ar\left[X \mid \Phi, \varsigma^2\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[X \mid \Phi, \varsigma^2\right]\right] + \mathbb{E}[X]^2 \\ &= \sigma_0^2 + \mathbb{V}ar[\Phi] + \mu_0^2 \\ &= \sigma_0^2 + \frac{\sigma_0^2}{k_0} + \mu_0^2 \\ &= \frac{k_0+1}{k_0}\sigma_0^2 + \mu_0^2. \end{aligned}$$

So the second parameter of the inverse-gamma distribution is

$$k_0\left(t_{02} - t_{01}\right)^2/2 = \sigma_0^2(k_0+1)/2 = \sigma_0^2(\gamma_0 - 1).$$

□

Corollary 4.5.5. *Consider a Bayesian model with* $X_j \,|\, \{\Phi = \varphi, \varsigma^2 = \sigma^2\} \sim \mathcal{N}(\varphi, \sigma^2)\ \forall j = 1, \ldots, k$ *and* $X_1, \ldots, X_k$ *conditional independent. Furthermore let*

$$\varsigma^2 \sim \mathcal{IG}\left(\gamma_0, (\gamma_0 - 1)\sigma_0^2\right)$$

with $\gamma_0 := (k_0+3)/2$ *and given* $\{\varsigma^2 = \sigma^2\}$

$$\Phi \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{k_0}\right).$$

Then the posterior distributions are

$$\varsigma \,|\, \{X_{1:k} = x_{1:k}\} \sim \mathcal{IG}\left(\gamma_k, \delta_k\right),$$

with

$$\gamma_k = \frac{k_0 + k + 3}{2} \quad \textit{and} \quad \delta_k = \frac{1}{2}\left((k_0+1)\sigma_0^2 + (k-1)s^2 + \frac{kk_0}{k_0+k}(\mu_0 - \bar{x})^2\right),$$

where $\bar{x} := 1/k \sum_{i=1}^{k} x_i$ *and* $s^2 := 1/(k-1)\sum_{i=1}^{k}(x_i - \bar{x})^2$ *and*

$$\Phi \,|\, \{X_{1:k} = x_{1:k}, \varsigma^2 = \sigma^2\} \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right),$$

with

$$\mu_k = \frac{k_0}{k_0+k}\mu_0 + \frac{k}{k_0+k}\bar{x} \quad \textit{and} \quad \sigma_k^2 = \frac{\sigma^2}{k_0+k}.$$

Proof. From Lemma 4.5.3 and Lemma 4.5.4 we directly get the parameters μ_k, σ_k^2 and γ_k. The parameter δ_k can be verified by the following calculation:
According to Lemma 4.5.3

$$\delta_k = \frac{1}{2}\left(k_0 t_{02} + k\overline{x^2} - \frac{1}{k_0+k}\left(k_0 t_{01} + k\bar{x}\right)^2\right).$$

Regarding Lemma 4.5.4 and $t_{01} = \mu_0$ and $t_{02} = (k_0+1)/k_0 \cdot \sigma_0^2 + \mu_0$, we obtain

$$\begin{aligned}
\delta_k &= \frac{1}{2}\left((k_0+1)\sigma_0^2 + k_0\mu_0^2 - \frac{k_0^2}{k_0+k}\varphi_0^2 - 2\frac{kk_0}{k_0+k}\mu_0\bar{x} - \frac{k^2}{k_0+k}\bar{x}^2 + k\overline{x^2}\right) \\
&= \frac{1}{2}\left((k_0+1)\sigma_0^2 + \frac{kk_0}{k_0+k}(\mu_0 - \bar{x})^2 - \frac{kk_0}{k_0+k}\bar{x}^2 - \frac{k^2}{k_0+k}\bar{x}^2 + k\overline{x^2}\right) \\
&= \frac{1}{2}\left((k_0+1)\sigma_0^2 + \frac{kk_0}{k_0+k}(\mu_0 - \bar{x})^2 - k\bar{x}^2 + k\overline{x^2}\right) \\
&= \frac{1}{2}\left((k_0+1)\sigma_0^2 + \frac{kk_0}{k_0+k}(\mu_0 - \bar{x})^2 + \sum_{i=1}^{k}(x_i - \bar{x})^2\right),
\end{aligned}$$

as

$$\begin{aligned}
\sum_{i=1}^{k}(x_i - \bar{x})^2 &= \sum_{i=1}^{k}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)^2 \\
&= k\overline{x^2} - 2k\bar{x}\bar{x} + k\bar{x}^2 \\
&= k\overline{x^2} - k\bar{x}^2.
\end{aligned}$$

□

Remark 4.5.6. *The posterior updating of the inverse gamma distribution can be interpreted nicely in the following manner. As the mean of* $\varsigma^2 \sim \mathcal{IG}(\gamma_k, \delta_k)$*-distribution,* $\gamma_k > 1$*, is (Remark B.12)*

$$\mathbb{E}\left[\varsigma^2\right] = \frac{\delta_k}{\gamma_k - 1},$$

the estimation of ς^2 is basically determined by δ_k (γ_k is just a scale factor). The update formula for δ_k is made of three terms, i.e.,

$$\delta_k = \frac{1}{2}\left((k_0+1)\sigma_0^2 + (k-1)s^2 + \frac{kk_0}{k_0+k}\left(\mu_0 - \bar{x}\right)^2\right).$$

The first term $(k_0+1)\sigma_0^2$ is just the weighted prior variance. The second term $(k-1)s^2$ is the weighted observed sample variance. To interpret the last term $kk_0/(k_0+k)(\mu_0-\bar{x})^2$ we think about μ_0 as the mean of k_0 prior observations. A large deviation from the next k observations is indicative for a high variance and should increase the posterior probability of a large σ^2.

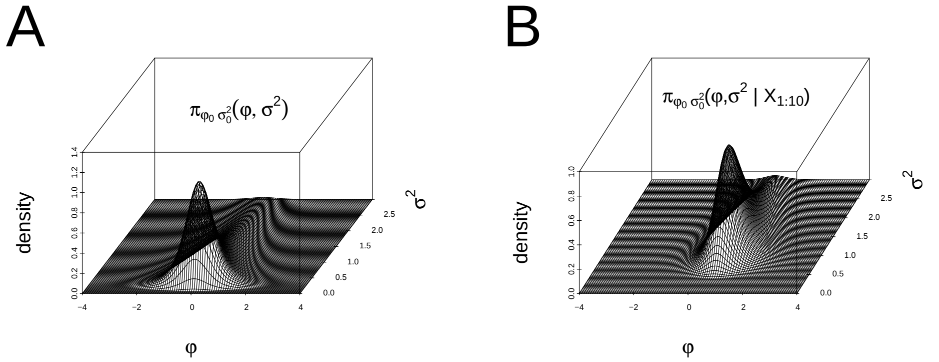


Figure 4.25: We choose prior parameter $\mu_0 = 0$ for the expected phase, $\sigma_0 = 1$ for the expected prior precision and $k_0 = 1$ as prior sample size. A. Joint prior distribution of (Φ, ς^2) for the chosen prior parameters. B. After observing $X_1, \dots, X_{10}$ with $X_i \sim \mathcal{N}(0.5, 1)$, we update our prior belief. As the true phase parameter is 0.5, the posterior distribution shifts to the right. Since we have additional information of $k = 10$ realizations, the distribution tightens in all directions.

4.5.3 Marginal distributions of Φ and ς^2

As ς^2 does not depend on Φ we directly see the marginal distribution of ς^2 in the joint distribution of ς^2 and Φ. An illustration of the updating process of the marginal distribution of ς^2 can be found in Figure 4.26 A.

Vice versa this is not as simple, as Φ depends on the value of ς^2. In Lemma 4.5.7 the marginal distribution of Φ is determined. Basically to get from $\Phi \mid \varsigma^2$ to Φ we only change the normal distribution to a t-distribution, taking the expectation of ς^2 into consideration. An illustration of the updating process of the marginal distribution of Φ can be found in Figure 4.27 A.

Lemma 4.5.7. *Let $\varsigma^2 \sim \mathcal{IG}(\gamma_k, \delta_k)$ and given $\{\varsigma^2 = \sigma^2\}$ let $\Phi \sim \mathcal{N}\left(\mu_k, \frac{\sigma^2}{k_0+k}\right)$ with γ_k, δ_k, μ_k as in Corollary 4.5.5 and $k \in \mathbb{N}$. Then the marginal distribution of Φ is*

$$\Phi \sim \mathcal{T}_{2\gamma_k}\left(\mu_k, \frac{\delta_k/\gamma_k}{k_0+k}\right),$$

where $\mathcal{T}_{2\gamma_k}$ is a t-distribution with $2\gamma_k$ degrees of freedom, cf. Definition B.13.

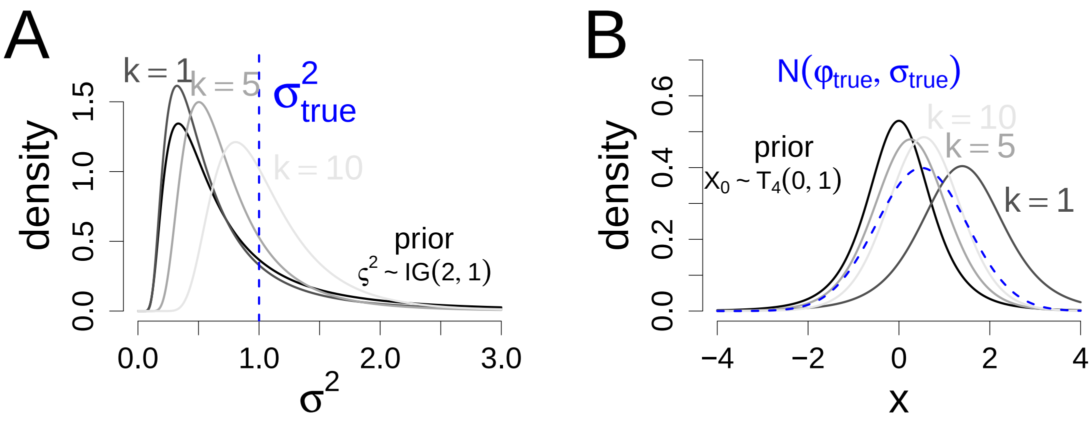


Figure 4.26: We start with prior phase parameter $\varphi_0 = 0$ and prior precision $\sigma_0 = 1$ and prior sample size $k_0 = 1$. A. Marginal update process of ς^2. According to the prior parameter choice the prior distribution is $\varsigma \sim \mathcal{IG}(2, 1)$ (black line). After observing $X_1, \ldots, X_k$ i.i.d. with $X_1 \sim \mathcal{N}(0.5, 1)$ for $k = 1, 5, 10$ the posterior distributions concentrate more and more at the true precision $\sigma = 1$ (blue dashed line). B. Predictive distribution. According to the prior parameters the prior distribution of X_0 is $X_0 \sim \mathcal{T}_4(0, 0.5)$ (black line), cf. Proposition 4.5.10. After observing $X_1, \ldots, X_k$, i.i.d. with $X_1 \sim \mathcal{N}(0.5, 1)$ for $k = 1, 5, 10$ the predictive distribution of $X_{k+1} \mid \{X_{1:k} = x_{1:k}\}$ gets closer to the true distribution $X_{k+1} \sim \mathcal{N}(0.5, 1)$ (blue dashed line).

Proof. Let $\pi_{\varphi,\sigma^2}(\cdot, \cdot)$ denote the joint prior distribution of Φ and ς. Integrating over σ^2 yields for the marginal distribution $\pi_\varphi(\cdot)$

$$\begin{aligned}\pi_\varphi(\varphi) &= \int_0^\infty \pi_{\varphi,\sigma^2}(\varphi, \sigma^2)\, d\sigma^2 \\ &\sim \int_0^\infty (\sigma^2)^{-1/2} \exp\left(-\frac{(\varphi - \mu_k)^2}{2\sigma^2/(k_0 + k)}\right) (\sigma^2)^{-(\gamma_k+1)} \exp\left(-\frac{\delta_k}{\sigma^2}\right) d\sigma^2 \\ &\sim \int_0^\infty (\sigma^2)^{-(\gamma_k+3/2)} \exp\left(-\frac{(k_0 + k)(\varphi - \mu_k)^2 + 2\delta_k}{2\sigma^2}\right) d\sigma^2,\end{aligned}$$

which corresponds to an unnormalized $\mathcal{IG}(\gamma_k + 1/2, 1/2((k_0 + k)(\varphi - \mu_k)^2 + 2\delta_k))$-distribution. As the integral equals the inverse normalization factor of the inverse-gamma distribution, the marginal distribution of φ is proportional to the inverse normalization factor, i.e.,

$$\begin{aligned}\pi_\varphi(\varphi) &\sim \frac{\Gamma\left(\gamma_k + 1/2\right)}{\left(\frac{(k_0+k)(\varphi-\mu_k)^2+2\delta_k}{2}\right)^{\gamma_k+1/2}} \\ &\sim \left(\frac{(k_0+k)(\varphi-\mu_k)^2+2\delta_k}{2}\right)^{-(\gamma_k+1/2)} \\ &\sim \left(1+\frac{(k_0+k)(\varphi-\mu_k)^2}{2\delta_k}\right)^{-(\gamma_k+1/2)} \\ &\sim \left(1+\frac{1}{2\gamma_k}\frac{(\varphi-\mu_k)^2}{\frac{\delta_k}{\gamma_k(k_0+k)}}\right)^{-(2\gamma_k+1)/2},\end{aligned}$$

which corresponds to a $\mathcal{T}_{2\gamma_k}\left(\mu_k, \frac{\delta_k/\gamma_k}{k_0+k}\right)$-distribution, cf. Definition B.13. □

Remark 4.5.8. *Since the variance of $X \sim \mathcal{T}_\nu(\mu, \sigma^2)$ is (Remark B.15)*

$$\mathbb{V}ar[X] = \frac{\nu}{\nu-2}\sigma^2,$$

we obtain according to Lemma 4.5.7

$$\mathbb{V}ar[\Phi \,|\, X_{1:k} = x_{1:k}] = \frac{2\gamma_k}{2(\gamma_k-1)}\frac{\delta_k/\gamma_k}{k_0+k} = \frac{\delta_k/(\gamma_k-1)}{k_0+k}.$$

As one might expect, regarding

$$\mathbb{E}[\varsigma \,|\, X_{1:k} = x_{1:k}] = \frac{\delta_k}{\gamma_k - 1},$$

the variance parameter of the t-distribution is chosen such

$$\mathbb{V}ar[\Phi \,|\, X_{1:k} = x_{1:k}] = \frac{\mathbb{E}[\varsigma^2 \,|\, X_{1:k} = x_{1:k}]}{k_0+k}.$$

Remark 4.5.9. *The difference of*

$$\mathcal{N}\left(\mu_k. \frac{\delta_k/(\gamma_k-1)}{k_0+k}\right) \quad \text{and} \quad \mathcal{T}_{2\gamma_k}\left(\mu_k, \frac{\delta_k/\gamma_k}{k_0+k}\right)$$

is negligible even for the first update process, see Figure 4.27 B. The prior distributions differ slightly.

4.5.4 Predictive distribution

To apply the BOCD to the setting of unknown phase and unknown precision, we still need to determine the predictive distribution of $X_{k+1} \,|\, \{X_{1:k} = x_{1:k}\}$, which is determined in Proposition 4.5.10. An illustration of the updating process of the predictive distribution can be found in Figure 4.26 B.

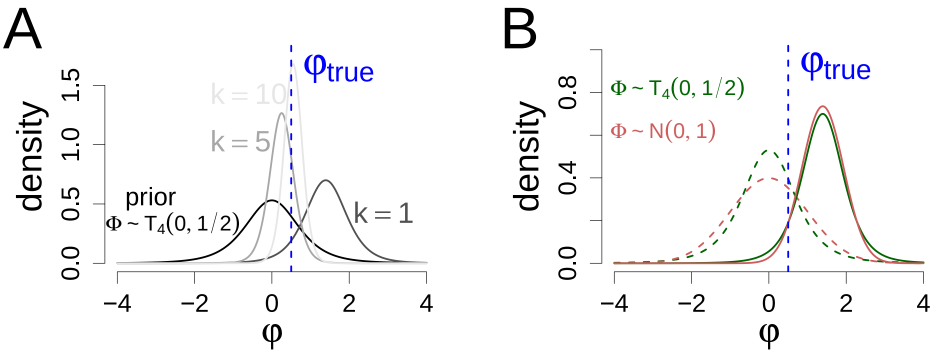


Figure 4.27: We start with prior phase parameter $\varphi_0 = 0$ and prior precision $\sigma_0 = 1$ and prior sample size $k_0 = 1$. A. Marginal update process of Φ^2. According to the prior parameter choice the prior distribution is $\Phi \sim \mathcal{T}(0, 1/4)$ (black line). After observing $X_1, \dots, X_k$, i.i.d. with $X_1 \sim \mathcal{N}(0.5, 1)$ for $k = 1, 5, 10$, the posterior distributions concentrate more and more at the true phase parameter $\varphi = 0.5$ (blue dashed line). B. Consider the same prior parameters and we observe the same realization $X_1, \dots, X_k$. What if we exchange the t-distribution (orange), by a normal distribution (green) for an approximate solution, see Remark 4.5.9? The prior distributions (dashed lines) differ slightly, even they have the same variance. But already after one realization, both posterior distributions are almost the same.

Proposition 4.5.10. *Consider a Bayesian model with* $X_j \,|\, \{\Phi = \varphi, \varsigma^2 = \sigma^2\} \sim \mathcal{N}(\varphi, \sigma^2)$ $\forall\, j = 1, \dots, k$ *and* $X_1, \dots, X_k$ *conditional independent. Furthermore, let*

$$\varsigma^2 \sim \mathcal{IG}\left(\gamma_0, (\gamma_0 - 1)\sigma_0^2\right)$$

with $\gamma_0 := (k_0 + 3)/2$ *and* $\sigma_0, k_0 > 0$. *Given* $\{\varsigma^2 = \sigma^2\}$ *let*

$$\Phi \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{k_0}\right).$$

Then the predictive distribution of $X_{k+1} \,|\, \{X_{1:k} = x_{1:k}\}$ *is*

$$X_{k+1} \,|\, \{X_{1:k} = x_{1:k}\} \sim \mathcal{T}_{2\gamma_k}\left(\mu_k, \frac{k_0 + k + 1}{k_0 + k} \frac{\delta_k}{\gamma_k}\right),$$

with

$$\gamma_k = \frac{k_0 + k + 3}{2} \quad \text{and} \quad \delta_k = \frac{1}{2}\left((k_0 + 1)\sigma_0^2 + (k - 1)s^2 + \frac{k k_0}{k_0 + k}(\mu_0 - \bar{x})^2\right)$$

and

$$\mu_k = \frac{k_0}{k_0 + k}\mu_0 + \frac{k}{k_0 + k}\bar{x} \quad \text{and} \quad \sigma_k^2 = \frac{\sigma^2}{k_0 + k},$$

where $\bar{x} := 1/k \sum_{i=1}^{k} x_i$ *and* $s^2 := 1/(k - 1) \sum_{i=1}^{k} (x_i - \bar{x})^2$.

Proof. With Corollary 4.5.5 we know the posterior distribution $\pi_{\varphi,\sigma^2}(\cdot \mid X_{1:k} = x_{1:k})$. Integrating over the possible values of φ and σ^2 and plugging in the posterior distribution $\pi_{\varphi,\sigma^2}(\cdot \mid X_{1:k} = x_{1:k})$ yields for the predictive distribution $p(\cdot \mid X_{1:k} = x_{1:k})$

$$
\begin{aligned}
p\left(x_{k+1} \mid X_{1:k} = x_{1:k}\right) &= \int_0^\infty \int_{-\infty}^\infty p\left(x_{k+1} \mid \Phi = \varphi, \varsigma^2 = \sigma^2\right) \pi_{\varphi,\sigma^2}(\varphi, \sigma^2 \mid X_{1:k} = x_{1:k})\, d\varphi\, d\sigma^2 \\
&= \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_{k+1} - \varphi)^2}{2\sigma^2}\right) \\
&\cdot \sqrt{\frac{k_0 + k}{2\pi\sigma^2}} \exp\left(-\frac{(\mu_k - \varphi)^2}{2\sigma^2/(k_0 + k)}\right) \frac{\delta_k^{\gamma_k}}{\Gamma(\gamma_k)} \left(\sigma^2\right)^{-\gamma_k - 1} \exp\left(-\frac{\delta_k}{\sigma^2}\right) d\varphi\, d\sigma^2 \\
&\sim \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma^2} \exp\left(-\frac{1}{2}\left(\underbrace{\left(\sigma^2\right)^{-1} (x_{k+1} - \varphi)^2 + \frac{k_0 + k}{\sigma^2}(\mu_k - \varphi)^2}_{\star}\right)\right) d\varphi \\
&\cdot \left(\sigma^2\right)^{-\gamma_k - 1} \exp\left(-\frac{\delta_k}{\sigma^2}\right) d\sigma^2,
\end{aligned}
$$

where we rewrite $\star$ as

$$
\star = \left(\varphi - \left(\frac{x_{k+1}}{k_0 + k + 1} + \frac{k_0 + k}{k_0 + k + 1}\mu_k\right)\right)^2 \frac{k_0 + k + 1}{\sigma^2} + (x_{k+1} - \mu_k)^2 \frac{k_0 + k}{k_0 + k + 1}\sigma^{-2}
$$

Using this identity we recognize that the inner integral is just an integrating over an unnormalized normal density and reduces to

$$
\int_{-\infty}^\infty \frac{1}{\sigma^2} \exp\left(-\frac{1}{2}(\star)\right) d\varphi = \sqrt{\frac{2\pi}{(k_0 + k + 1)\sigma^2}} \exp\left(\frac{1}{2}\frac{k_0 + k}{k_0 + k + 1}\sigma^{-2}(x_{k+1} - \mu_k)^2\right).
$$

Overall this yields for the predictive distribution

$$
p\left(x_{k+1} \mid X_{1:k} = x_{1:k}\right) \sim \int_0^\infty \left(\sigma^2\right)^{-\gamma_k - 3/2} \exp\left(-\frac{\delta_k + \frac{1}{2}\frac{k_0+k}{k_0+k+1}(x_{k+1} - \mu_k)^2}{\sigma^2}\right) d\sigma^2,
$$

which is an unnormalized inverse-gamma distribution, thus

$$
\begin{aligned}
p\left(x_{k+1} \mid X_{1:k} = x_{1:k}\right) &\sim \left(\delta_k + \frac{1}{2}\frac{k_0 + k}{k_0 + k + 1}(x_{k+1} - \mu_k)^2\right)^{-(\gamma_k + 1/2)} \\
&\sim \left(1 + \frac{1}{2}\frac{k_0 + k}{k_0 + k + 1}\frac{1}{\delta_k}(x_{k+1} - \mu_k)^2\right)^{-\frac{2\gamma_k + 1}{2}} \\
&\sim \left(1 + \frac{1}{2\gamma_k}\frac{\gamma_k(k_0 + k)}{k_0 + k + 1}\frac{1}{\delta_k}(x_{k+1} - \mu_k)^2\right)^{-\frac{2\gamma_k + 1}{2}},
\end{aligned}
$$

corresponding to a t-distribution with $2\gamma_k$ degrees of freedom, location parameter μ_k and scale parameter $\frac{k_0+k+1}{k_0+k}\frac{\delta_k}{\gamma_k}$. □

Remark 4.5.11. *The predictive distribution of $X_{k+1} \mid \{X_{1:k} = x_{1:k}\}$ (Proposition 4.5.10) and the marginal posterior distribution of $\Phi \mid \{X_{1:k} = x_{1:k}\}$ (Lemma 4.5.7) are both t-distributions* *with $2\gamma_k$ degrees of freedom and mean μ_k, but the variance of the predictive distributions is $\delta_k/(\gamma_k - 1)$ higher, i.e.,*

$$\begin{aligned}\mathbb{V}ar[X_{k+1} \,|\, X_{1:k} = x_{1:k}] &= \frac{2\gamma_k}{2\gamma_k - 2}\frac{k_0+k+1}{k_0+k}\frac{\delta_k}{\gamma_k} \\ &= \frac{k_0+k+1}{k_0+k}\frac{\delta_k}{\gamma_k - 1} \\ &= \frac{1}{k_0+k}\frac{\delta_k}{\gamma_k-1} + \frac{\delta_k}{\gamma_k - 1} \\ &= \mathbb{V}ar[\Phi \,|\, X_{1:k} = x_{1:k}] + \mathbb{E}[\varsigma^2 \,|\, X_{1:k} = x_{1:k}].\end{aligned}$$

This equation also follows directly from the law of total variance, which says for $X \sim \mathcal{N}(\Phi, \varsigma^2)$

$$\begin{aligned}\mathbb{V}ar[X] &= \mathbb{V}ar\left[\mathbb{E}_{\{\Phi,\varsigma^2\}}[X]\right] + \mathbb{E}\left[\mathbb{V}ar_{\{\Phi,\varsigma^2\}}[X]\right] \\ &= \mathbb{V}ar\left[\Phi\right] + \mathbb{E}\left[\varsigma^2\right].\end{aligned}$$

Remark 4.5.12. *Applying Theorem 3.4.20 we directly know the posterior expectation of the sufficient statistic $t(X)$, since we have an exponential family distribution and a standard conjugate prior, thus posterior linearity holds. As we chose*

$$t_{01} = \mu_0 \quad \textit{and} \quad t_{02} = \frac{k_0+1}{k_0}\sigma_0^2 + \mu_0^2,$$

we obtain for $t(x) = \big(x, x^2\big)$

$$\mathbb{E}\left[t(X_{k+1}) \,|\, X_{1:k} = x_{1:k}\right] = \begin{pmatrix} \frac{k_0\mu_0 + k\bar{x}}{k_0+k} \\ \frac{(k_0+1)\sigma_0^2 + k_0\mu_0^2 + k\bar{x^2}}{k_0+k}\end{pmatrix}.$$

That can be also verified by Proposition 4.5.10, where $\mathbb{E}\left[X_{k+1} \,|\, X_{1:k} = x_{1:k}\right]$ is obvious identically. Since

$$X_{k+1} \,|\, \{X_{1:k} = x_{1:k}\} \sim \mathcal{T}_{2\gamma_k}\left(\mu_k, \frac{k_0+k+1}{k_0+k}\frac{\delta_k}{\gamma_k}\right),$$

and thus

$$\mathbb{V}ar\left[X_{k+1} \,|\, \{X_{1:k} = x_{1:k}\}]\right) = \frac{2\gamma_k}{2\gamma_k - 2}\frac{k_0+k+1}{k_0+k}\frac{\delta_k}{\gamma_k},$$

we obtain for the second part

$$\begin{aligned}\mathbb{E}\left[X_{k+1}^2 \,|\, X_{1:k} = x_{1:k}\right] &= \mathbb{V}ar\left[X_{k+1} \,|\, X_{1:k} = x_{1:k}\right] + \left(\mathbb{E}\left[X_{k+1} \,|\, X_{1:k} = x_{1:k}\right]\right)^2 \\ &= \frac{(k_0+1)\sigma_0^2 + (k-1)s^2 + \frac{kk_0}{k_0+k}\left(\mu_0 - \bar{x}\right)^2}{k_0+k} + \left(\frac{k_0}{k_0+k}\mu_0 + \frac{k}{k_0+k}\bar{x}\right)^2 \\ &= \frac{(k_0+1)\sigma_0^2 + k\bar{x^2} - k\bar{x}^2 + \frac{k_0k}{k_0+k}\left(\mu_0 - \bar{x}\right)^2 + \frac{k_0^2}{k_0+k}\mu_0^2 + 2\frac{k_0k}{k_0+k}\mu_0\bar{x} + \frac{k^2}{k_0+k}\bar{x}^2}{k_0+k} \\ &= \frac{(k_0+1)\sigma_0^2 + k\bar{x^2} - k\bar{x}^2 + \frac{k_0(k+k_0)}{k_0+k}\mu_0^2 + \frac{k(k+k_0)}{k_0+k}\bar{x}^2}{k_0+k} \\ &= \frac{(k_0+1)\sigma_0^2 + k_0\mu_0^2 + k\bar{x^2}}{k_0+k}.\end{aligned}$$

4.5.5 Application of BOCD

With the results of the previous sections we are able to apply the BOCD to the case that rate Λ, phase Φ and spike precision ς^2 can change at some point in time. The change point model assumes that in oscillation cycle k

$$N_k \sim Pois(\Lambda_{A_k})$$

spikes occur and given $\{N_k = n_k\}$ we choose spike times

$$X_1^{(k)}, \dots, X_{n_k}^{(k)} \sim \mathcal{N}(\Phi_{A_k}, \varsigma^2_{A_k}),$$

where rate, phase and precision parameter can not change in an oscillation cycle, but between adjacent cycles.

We observe a sequence of cycles and need to decide for change points, i.e., we want to find the time points, where the parameters Λ, Φ and ς^2 change. Again we assume that changes in the variables occur simultaneously (cf. Section 4.1). As theoretically the new parameters are chosen according to their prior distribution, it is possible that for example the rate before and after the change point remains almost the same.

By applying the BOCD we assume that if a change point occurs, the new parameters are chosen according to our prior distributions, i.e.,

$$\Lambda \sim \mathcal{G}amma(\alpha_0, \beta_0) \quad \text{and} \quad \varsigma^2 \sim \mathcal{IG}(\gamma_0, (\gamma_0 - 1)\sigma_0^2)$$

and given $\{\varsigma^2 = \sigma^2\}$

$$\Phi \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{k_0}\right).$$

In the following simulations we use the same prior parameters as in Section 4.3.3

$$\alpha_0 = 3,\ \beta_0 = 1; \quad k_0 = 1\,(\Rightarrow \gamma_0 = 2),\ \sigma_0 = 1; \quad \mu_0 = 0.$$

We apply the BOCD and assume that change points occur with a constant but unknown probability H, cf. Section 4.2.2, and use an uniform change point prior ($a_0 = b_0 = 1$).

To explore how changes in the precision ς^2 affects the change detection in Φ, we additional consider an reduced model (pure phase analysis), where we only consider changes in Φ and view $\varsigma^2 = \sigma^2$ as fixed, i.e., given $\{N_k = n_k\}$ we assume in oscillation cycle k

$$X_1^{(k)}, \dots, X_{n_k}^{(k)} \sim \mathcal{N}(\Phi, \sigma^2).$$

Again we apply the BOCD to search for change points, but here only for changes in rate Λ and phase Φ. Here we choose the prior parameter $\tau_0 = 2$ (see Section 4.3.3.2).

For the various types of analysis we use the following short cuts:

1. Only λ: Pure rate analysis. We only use the number of spikes in each oscillation cycle, so we apply the BOCD on Λ.

2. Only φ: Pure phase analysis. We consider $\sigma^2 = 1$ as fixed and only use the spike times $X_j^{(k)}$ in each oscillation cycle and do not consider the different spike numbers. So we apply the BOCD on Φ.

3. $\varphi + \sigma$: Phase and precision analysis. We only use the spike times $X_j^{(k)}$ in each oscillation cycle, but consider simultaneous changes in phase and precision. So we apply the BOCD on Φ and ς^2.

4. $\lambda + \varphi + \sigma$: Rate, phase and precision analysis. We use the number of spikes and the spike times $X_j^{(k)}$ in each oscillation cycle and apply the BOCD on Λ, Φ and ς^2.

First change point setting Changes only in the precision ς^2, see Figure 4.28 A and B. Here we consider spike trains consisting of $K = 10, 20, \ldots, 100$ cycles, where no rate and phase change occur, but the spike precision changes, i.e.,

$$N_k \sim Pois(4), \quad k = 1, \ldots, K$$

and for each $j = 1, \ldots, N_k$

$$X_j^{(k)} \sim \mathcal{N}(0,1), \quad k = 1, \ldots, K/2$$

and

$$X_j^{(k)} \sim \mathcal{N}\left(0, 2^2\right), \quad k = K/2+1, \ldots, K.$$

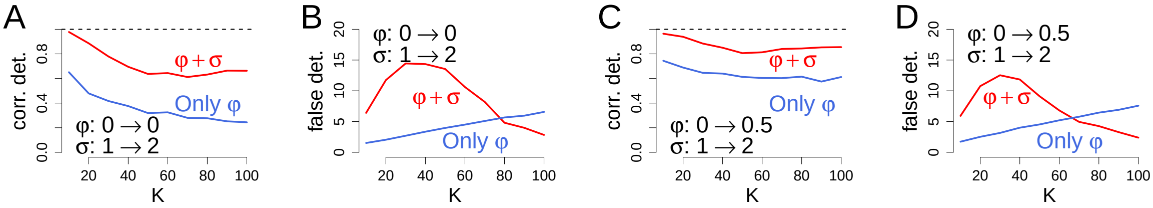


Figure 4.28: Evaluation of BOCD using pure phase analysis or phase and precision analysis in the one change point setting with a change point in the middle. We choose prior parameters $k_0 = 1$, $\sigma_0 = 1$, $\mu_0 = 0$ and $\tau_0 = 2$. The precision changes from $\sigma = 1$ to $\sigma = 2$, the phase remains constant (first change point setting)(A,B) or changes from 0 to 0.5 (second change point setting) (C,D). Average number of correctly detected change points (A,C) and average number of falsely detected change points (B,D) for sequences of length $K = 10$ to $K = 100$, 1000 simulations per data point. Pure phase analysis ($\sigma = 1$) is shown in blue and phase and precision analysis in red.

If we apply the BOCD with a pure phase analysis, so we consider the spike precision as fixed with $\sigma = 1$ and only look for changes in phase Φ, we detect the correct change point at $K/2$ in some simulations, but the number of correctly detected change points decreases with the number of cycles K, see blue line in Figure 4.28 A. More interestingly the number of falsely detected change points increases with the number of cycles (Figure 4.28 B, blue line). In case of a phase and precision analysis, so we regard also changes in the spike precision ς^2, we detect the correct change point more often (Figure 4.28 A, red line) and more importantly the number of false detections decreases with an increasing number of cycles K (Figure 4.28 B, red line).

Second change point setting Changes in phase Φ and precision ς^2 (Figure 4.28 C, D). Here we consider spike trains consisting of $K = 10, 20, \ldots, 100$ cycles, where no rate change occurs, but phase and spike precision change, i.e.,

$$N_k \sim Pois(4), \quad k = 1, \ldots, K$$

and for each $j = 1, \ldots, N_k$

$$X_j^{(k)} \sim \mathcal{N}(0, 1), \quad k = 1, \ldots, K/2$$

and

$$X_j^{(k)} \sim \mathcal{N}\left(0.5, 2^2\right), \quad k = K/2 + 1, \ldots, K.$$

In case of a pure phase analysis, i.e., we consider the spike precision as fixed with $\sigma = 1$, we detect the correct change point at $K/2$ only in about $2/3$ of cases, see blue line in Figure 4.28 C. But again the number of falsely detected change points increases with the number of cycles (Figure 4.28 D, blue line).

However in case of a phase and precision analysis, we almost ever detect the change point correctly (Figure 4.28 C, red line) and simultaneously the number of false detections decreases with an increasing number of cycles K (Figure 4.28 D, red line).

In summary, it is very important to consider the spike precision as variable, if we have the opinion that the spike precision can change. Otherwise, the change detection in the phase can not work reliable.

Third change point setting Changes in rate Λ, phase Φ and precision ς^2 (Figure 4.29). Here we consider spike trains consisting of $K = 10, 20, \ldots, 200$ cycles, where rate, phase and spike precision change simultaneously, i.e.,

$$N_k \sim Pois(\lambda), \quad k = 1, \ldots, K/2$$

and

$$N_k \sim Pois(2\lambda), \quad k = K/2 + 1, \ldots, K$$

and for each $j = 1, \ldots, N_k$

$$X_j^{(k)} \sim \mathcal{N}(0, 1), \quad k = 1, \ldots, K/2$$

and

$$X_j^{(k)} \sim \mathcal{N}\left(0.5, 2^2\right), \quad k = K/2 + 1, \ldots, K.$$

The case of a low spike number ($\lambda = 1$) is shown in Figure 4.29 A and B, for a high spike number ($\lambda = 2$) see Figure 4.29 C and D.

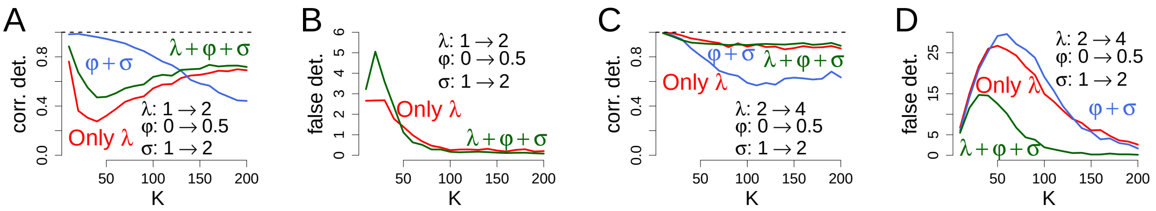


Figure 4.29: Evaluation of BOCD using pure rate or phase and precision or rate, phase and precision analysis in the one change point setting with a change point in the middle (third change point setting). We choose prior parameters $\alpha_0 = 3$, $\beta_0 = 1$, $k_0 = 1$, $\sigma_0 = 1$, $\mu_0 = 0$ and $\tau_0 = 2$. The precision changes from 1 to 2, the phase changes from 0 to 0.5, the rate changes from 1 to 2 (low spike number) (A,B) or changes from 2 to 4 (high spike number) (C,D). Average number of correctly detected change points (A,C) and average number of falsely detected change points (B,D) for sequences of length $K = 10$ to $K = 200$, 1000 simulations per data point. Pure rate analysis is shown in red, phase and precision analysis in blue and a rate, phase and precision analysis is shown in green.

Results - Low Spike number, Figure 4.29 A and B:

In case of a low spike number a pure rate analysis (red line) needs at least $K = 100$ cycles to detect in about 50% of cases the change point correctly (panel A). Due to the small variance of the Poisson distribution the change detection is quite robust and only some false detections arise (panel B).

Using phase and spike precision ($\varphi + \sigma$) analysis results in a high number of falsely detected change points (panel B, the blue line lies beyond the plot window and is not shown). Due to the high number of false detections also the correct change point is detected (panel A), but the number of correctly detected change point decreases with the number of cycles ($K = 200$), as also the number of false detections reduces. Due to the low spike number the detection ability of phase and precision worsens compared to Figure 4.28 ($\lambda = 4$), as the mean spike time is less accurate.

Nevertheless, the pure rate analysis can be improved by also using the information contained in the phase and spike precision. The $\lambda + \varphi + \sigma$ analysis (green line) results in a higher number of correctly detected change points (panel A) and the number of false detections is only increased for $K < 50$ cycles (panel B).

Results - High Spike number 4.29 C and D:

In case of a high spike number a pure rate analysis (red line) detects the correct change point almost always, independent of the number of cycles K (panel C). Due to the higher variance of the Poisson distribution (compared to the case of low spike number) the change detection is less robust and a high number of cycles ($K \approx 200$) is needed for a small number of falsely detected change points (panel D).

Here a phase and spike precision analysis ($\varphi + \sigma$) results in an increased number of correctly detected change points (panel C, blue line) compared to the case of low spike number (panel

A, blue line) and the number of fales detections is significantly increased and similar to a pure rate analysis (panel D). A higher spike number increases the validity of the mean spike time. More interestingly, if we combine the information contained in the number of spikes, the phase and spike precision (green line), the number of correctly detected change points is almost identical to a pure rate analysis (panel C). But a $\lambda + \varphi + \sigma$ analysis reduces the number of false detections strikingly, and almost no false detections occur for at least $K = 100$ cycles (panel D) and the result is a very robust change detection.

In summary also in case of an unknown spike precision imprecise phases, compared to a pure rate analysis, can increase the number of correctly detected change points and more importantly decrease the number of false detections especially in case of high spike numbers.

Chapter 5

Empirical neurons

In this section we apply our theoretical results on the detection probability and our algorithms for change point detection to a setting of empirical neurons as reported in Havenith et al. (2011). The authors recorded eight neurons in response to 12 stimuli, which were drifting sinusoidal gratings of which the drifting direction rotated in steps of 30° (Figure 5.14 A).

First we present the rate and phase parameters of each empirical neuron, see Section 5.1. In Section 5.2 we concentrate on the stimulus encoding. We consider the detection probability of each stimulus and each neuron individually and observe that more than half of the neurons are 'rate neurons' that decide mostly for one stimulus, if a spike occurs, cf. Section 5.2.1. Based on the results of Section 2.1.2.3 we determine parameters of theoretical neurons that maximize the detection probability for the same rate and phase parameter ranges. Dependent on the parameter range of each empirical neuron the detection probability can be increased up to a third, but some neurons encode stimuli near optimal. However, the increase in the detection probability by using rate and phase compared to a pure rate analysis is quite small, for most of the single empirical neurons and the consideration of all empirical neurons simultaneously.

In Section 5.2.2 we draw on the results of Section 2.1.3 and determine theoretical neurons that minimize the detection error. Similar to the detection probability the detection error can be decreased up to a third. However, even if the theoretical neurons with maximal detection probability and the theoretical neurons with minimal detection error can have quite different parameter structure, the resulting detection probability and detection error are almost identical, which motivates not too focus to much on these optimizations. Therefore, we focus on the empirical neurons and the probability $p^{(\delta)}$ of falsely detecting an incorrect stimulus as a function of the distance δ between the correct and incorrect stimulus. Already considering the single neurons we observe that false decisions rather occur between stimuli with small distances. More interestingly considering all empirical neurons simultaneously yields almost only false decisions between stimuli with a distance of $\delta = 1$.

In Section 5.3 we concentrate on the change point detection (a detected change point is correct, if the distance to a true change point is at most three). First we explain, how we determine the predictive distribution, as we are now in a discrete change point setting with discrete rate and phase priors, cf. Section 5.3.1. Furthermore, we note that the parameters of a theoretic neuron that maximizes the detection probability is not necessarily optimal with respect to the change point detection.

In Section 5.3.2 we compare one empirical neuron and its theoretical optimal neuron with respect to the performance in the change point detection. Further we quantify the improvement in the change point detection by the bivariate analysis compared to a pure rate analysis. The improvement is clearer in the theoretical neuron and increases with higher rate parameters, for example all rate parameters multiplied by four, which is equivalent to four times the same responding neuron.

In Section 5.3.3 we consider all empirical neurons simultaneously to detect change points. We consider the change point detection dependent on the distance of changed stimuli, as most changes between stimuli can be detected very surely and change detection is only difficult for stimuli with a distance of $\delta = 1$. We observe that the phase increases the probability of correctly detecting a change point especially in case of highly similar stimuli and reduces the probability of falsely detecting a change point.

5.1 Rate and phase parameters of the empirical neurons

As described in Appendix A we use Figure 5 of Havenith et al. (2011) to roughly read off the measured number of spikes and the relative firing times:

$$
\begin{aligned}
\begin{pmatrix} \boldsymbol{\lambda}^{(1)} \\ \boldsymbol{\varphi}^{(1)} \end{pmatrix} &= \begin{pmatrix} 7.5 & 5 & 0 & 5 & 12.5 & 7.5 & 7.5 & 9 & 7.5 & 5 & 5 & 5 \\ 2 & 2 & 2 & 0 & -1 & 1 & -2 & -2 & -2 & -1 & 1 & -2 \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{\lambda}^{(2)} \\ \boldsymbol{\varphi}^{(2)} \end{pmatrix} &= \begin{pmatrix} 12.5 & 0 & 0 & 0 & 0 & 8 & 42 & 42 & 17 & 8 & 17 & 25 \\ -2 & -4 & -2 & -2 & -3 & -3 & -5 & -7 & -6 & -4 & -3 & -2 \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{\lambda}^{(3)} \\ \boldsymbol{\varphi}^{(3)} \end{pmatrix} &= \begin{pmatrix} 4 & 4 & 17 & 25 & 8 & 4 & 4 & 4 & 8 & 12 & 8 & 4 \\ -3 & 0 & 0 & 1 & 0 & -3 & 1 & 3 & 3 & 1 & -3 & -1 \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{\lambda}^{(4)} \\ \boldsymbol{\varphi}^{(4)} \end{pmatrix} &= \begin{pmatrix} 0 & 4 & 8 & 0 & 0 & 0 & 4 & 12 & 25 & 4 & 4 & 0 \\ -2 & -2.5 & -3 & -1.5 & 0 & -1 & 1 & -2 & -2 & -1 & -2 & 0 \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{\lambda}^{(5)} \\ \boldsymbol{\varphi}^{(5)} \end{pmatrix} &= \begin{pmatrix} 13 & 33 & 33 & 7 & 0 & 0 & 7 & 14 & 7 & 0 & 0 & 7 \\ 3 & 2 & 3.5 & 6 & 6 & 4 & 2 & 1 & 1 & 3 & 1 & 0 \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{\lambda}^{(6)} \\ \boldsymbol{\varphi}^{(6)} \end{pmatrix} &= \begin{pmatrix} 0 & 0 & 0 & 0 & 20 & 40 & 60 & 60 & 40 & 10 & 0 & 0 \\ 0 & 2 & 2 & -1 & 0 & 0 & 2 & 2 & 2 & 3 & 3 & 2 \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{\lambda}^{(7)} \\ \boldsymbol{\varphi}^{(7)} \end{pmatrix} &= \begin{pmatrix} 7.5 & 10 & 5 & 2.5 & 2.5 & 2.5 & 7.5 & 12.5 & 7 & 2.5 & 2.5 & 2.5 \\ 3 & 1 & 2 & 2 & -2 & -2 & -2.5 & -3.5 & -2 & 1 & 2 & 2 \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{\lambda}^{(8)} \\ \boldsymbol{\varphi}^{(8)} \end{pmatrix} &= \begin{pmatrix} 7.5 & 2.5 & 2.5 & 2.5 & 5 & 5 & 2.5 & 2.5 & 5 & 7.5 & 10 & 10 \\ 0 & 0 & -3 & -3 & 0 & 3 & 3 & 3 & 2 & -2 & 0 & 3 \end{pmatrix}
\end{aligned}
$$

For comparison with the theoretical results, the rate parameters of the empirical neuron are rescaled to measure the number of spikes per oscillation cycle, and the phase parameters are divided by $\hat{\sigma} \approx 6$ to obtain an approximate standard deviation of 1 as used in the theoretical considerations (for details see Appendix A).

The notation of rate and phase parameters of each neurons is analogous to Section 2.2. Additionally, let $\lambda_M^{(m)}$ denote the maximal rate parameter and $\varphi_M^{(m)}$ the maximal phase difference of neuron $m = 1, \ldots, 8$.

The observed rate and phase parameters of each stimulus individually for every neuron are illustrated in Figure 5.1. Note that we now have negative phase values as the firing times between neurons matter. The maximal phase difference within a single neuron is approximately between 0.6 and 1, the maximal rate is between 0.6 and 3.5. In Figure 5.1 we also show the detection probability p_D calculated for each neuron separately that ranges between 0.12 and 0.207. If we would randomly decide for a stimulus, we would obtain a detection probability of $1/12 \approx 0.08$, what questions, if we should take neurons $1, 3, 4, 7$ and 8 seriously. Further only neuron 2 and 6 (maybe 5) have considerable 'high' rate parameters (above 1) in more than one stimulus to distinguish more than one stimulus with the rate. For more details see the next section.

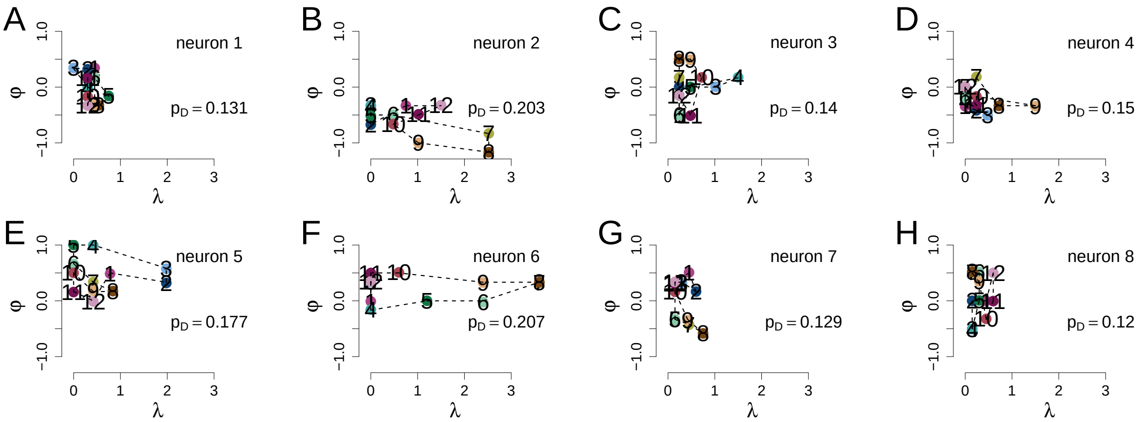


Figure 5.1: Scaled rate and phase parameters of the eight empirical neurons reported in Havenith et al. (2011). The number in each dot represents the stimulus in the circular order, cf. Figure 5.8 A, i.e., stimulus 5 has neighbors 4 and 6. The dashed line connects neighbored stimuli.

5.2 Stimulus encoding

In the following sections we draw on results obtained of Section 2. Especially to calculate the detection probability for one neuron we use Lemma 2.1.3 and for multiple neurons we determine the detection probability by simulations with Lemma 2.2.2. Explanations concerning the detection error or the average probability to misclassify two stimuli with a distance of δ can be found in Section 2.1.3.

5.2.1 Analysis of the detection probability

The main take away of this Section is that the empirical neurons code information close to optimal based on only rate, but only neuron 2 is close to an optimal rate and phase code. Considering all neurons simultaneously a rate and phase analysis can not increase the detection probability strikingly compared to a pure rate analysis.

In Figure 4.2 we plug in the rate and phase parameters of Figure 5.1 and determine the detection probability $p_s^{(m)}$ for each stimulus $s = 1, \ldots, 12$ separately for each neuron $m = 1, \ldots, 8$. Thereby we distinguish if we only use the rate (red), only the phase (blue) or rate and phase (green). First, we notice that 'only rate' and 'rate and phase' results in similar detection probabilities. Second, 'only phase' results in almost the same detection probabilities for all stimuli. Third, neurons $1, 3, 4, 7, 8$ are basically binary in the rate, as either stimuli with

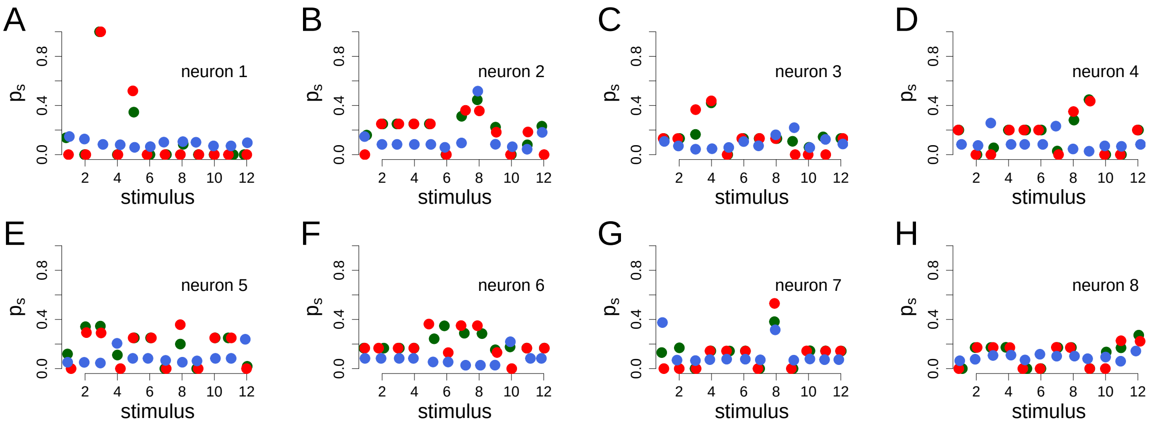


Figure 5.2: Detection probabilities $p_s^{(m)}$ for all $s = 1, \ldots, 12$ stimuli and $m = 1, \ldots, 8$ empirical neurons. A pure rate analysis is shown in red, pure phase analysis in blue and a rate and phase analysis in green.

minimal rate parameters are detected if no spike occurs or otherwise stimuli with maximal rate parameters are detected. For example let us consider neuron 1: The minimal rate of zero is assigned to stimulus 3, thus this stimulus is detected correctly with $p_3^{(1)} = 1$. The maximal rate is assigned to stimulus 5 and is less than one, stimulus 5 is always detected, if a spike occurs. All other stimuli are never detected in case of pure rate analysis. The same can be seen for neuron 8, where stimuli $2, 3, 4, 7, 8$ have the same minimal rate and stimuli 11 and 12 have the same maximal rate. All other stimuli are again never detected.

In Figure 5.3 we illustrate the rate and phase parameters that maximize the detection probability using the same maximal rate $\lambda_M^{(m)}$ and maximal phase range $\varphi_M^{(m)}$ as the empirical neuron $m = 1, \ldots, 8$. In black we note the detection probabilities of the theoretical neurons and in green the percentage, the theoretical neurons increase the detection probability in comparison to the empirical neurons. The maximal increase is about one third. Of special interest is neuron 2, which shows a relatively large range of rate and phase parameters across stimuli and shows a close-to-optimal behavior (increase only about 8.5, even we consider all stimuli. For more details to neuron 2 see Section 5.3.2.)

In Figure 5.4 A we can observe that the empirical neurons (except neuron 3) have at least the same detection probability (green dots, rate and phase) as a theoretic optimal neuron using only rate (red dots, note that this is also the optimal rate code due to the small rates). The detection probability of a theoretic optimal neuron using rate and phase is shown in black, which demonstrates that for theoretic neurons the detection probability can be strikingly increased by rate and phase analysis compared to a pure rate analysis.

However, for the empirical neurons only in case of neuron 2 the detection probability is strikingly increased by the rate and phase analysis (green) compared to only rate (violet, Figure 5.4 B). Interestingly, in case of pure rate analysis the empirical neurons (violet) have almost the same detection probability as a theoretic optimal neuron (red).

But Figure 5.4 C points up that this is no big deal, as neurons that only emit spikes for

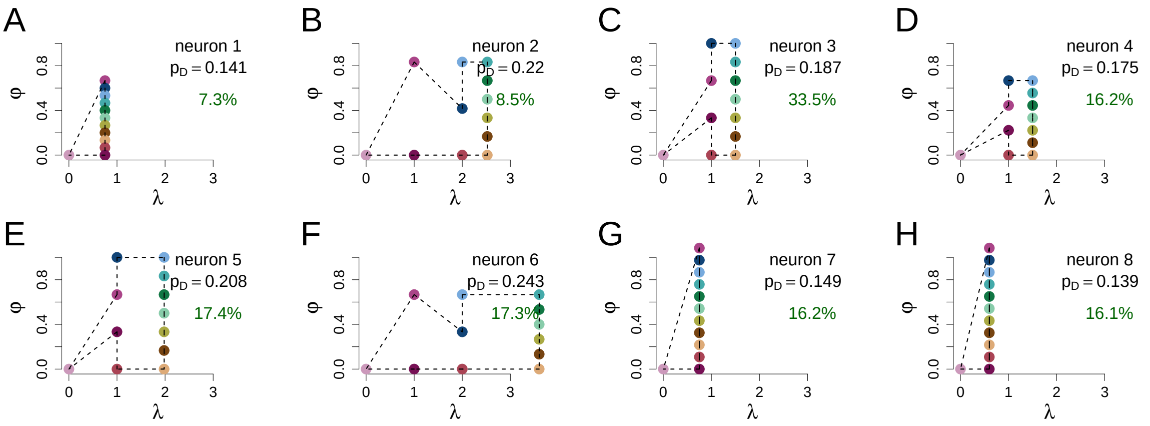


Figure 5.3: Theoretical optimal (detection probability) rate and phase parameters for the rate ($\lambda_M^{(m)}$) and phase ($\varphi_M^{(m)}$) parameter range of each empirical neuron $m = 1, \ldots, 8$. The detection probability of the theoretical neuron is shown in black and the increase compared to the empirical neuron is written in green.

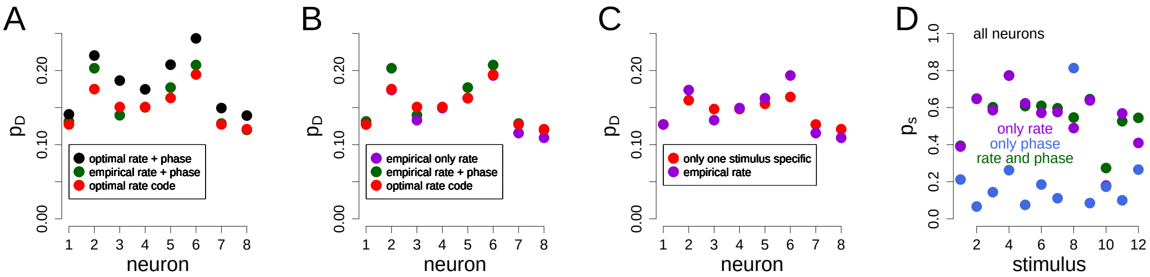


Figure 5.4: A-C. Detection probability p_D separately for single neurons. A and B. Theoretical optimal neurons using only rate (red) and rate and phase (black). The empirical neurons using only rate are shown in violet, using rate and phase in green. C. In red is shown a rate neuron that has maximal rate $\lambda_M^{(m)}$ for one stimulus and zero rate else. D. We consider all empirical neurons simultaneously and determine the detection probability separately for each stimulus. A pure rate analysis is shown in violet, pure phase analysis in blue and rate and phase analysis in green.

one stimulus (red) are comparable to the empirical neurons considering only the rate (violet) (except neuron 6).

Finally, we consider all neurons simultaneously and determine the detection probabilities p_s of each stimulus $s = 1, \ldots, 12$ by simulation, cf. Lemma 2.2.2. First we note that the global detection probability using 'rate and phase' is about $p_D \approx 0.56$, while only rate yields $p_D \approx 0.54$. This is supported by Figure 5.4, where only in case of stimulus 10 and 12 the detection probabilities using rate and phase (green) are increased compared to a pure rate analysis (violet). This is not surprising, as we have already seen theoretically in Section 2.2 that due to the efficient binary coding imprecise phases only provide additional information for $S > 2^M$ stimuli.

5.2.2 Distance of stimuli and false decisions

Here we apply our insights of Section 2.1.3 and compare the empirical neurons to theoretic neurons that minimize the detection error e_D. The main outcome is that the empirical neurons try to avoid false decisions to very different stimuli, already on a single neuron basis and more interestingly the set of eight neurons can correctly identify the correct stimulus with a precision of about ±30°.

In Figure 5.5 we present the rate and phase parameters which we obtained numerically by minimizing the detection error e_D. Analog to the detection probability the detection error can be decreased theoretically up to one third.

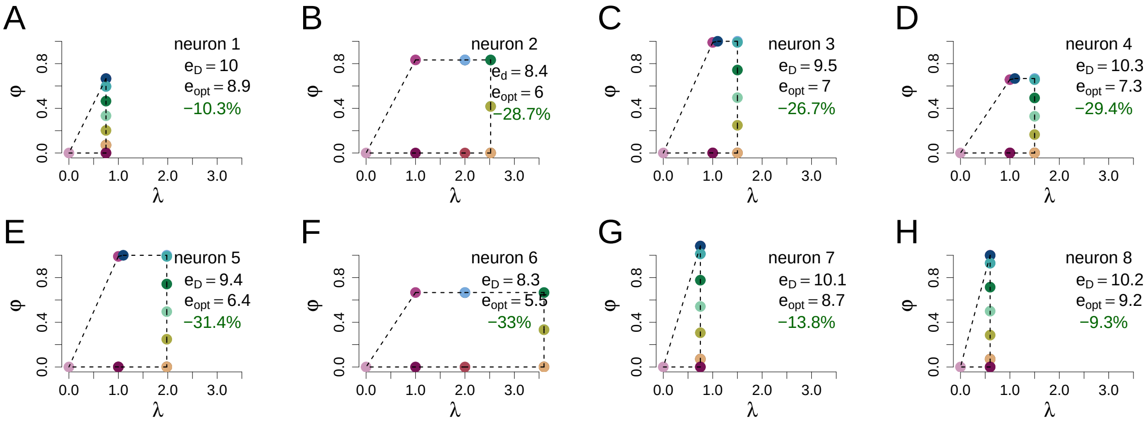


Figure 5.5: Theoretical optimal (detection error) rate and phase parameters for the rate $(\lambda_M^{(m)})$ and phase $(\varphi_M^{(m)})$ parameter range of each empirical neuron $m = 1, \dots, 8$. The detection error of the empirical neurons is denoted by e_D, of the theoretical optimal neurons by e_{opt}. The decrease compared to the empirical neuron is written in green.

However, as Figure 5.6 suggests it seems not very plausible to distinguish between the theoretically optimal parameter set that maximizes the detection probability (red) and the optimal parameter set that minimizes the detection error (black), as both result in almost the same detection probability (panel B) or detection error (panel A).

Therefore, we further consider the empirical neurons and investigate those cases with incorrect detections and show the results as a function of the 'distance' between stimuli as follows. The stimuli in Havenith et al. (2011) were gratings drifting in twelve different directions, such that the distance δ between two stimuli can be determined naturally as a function of the drifting angle (step size 30°, see Figure 5.8 A).

Figure 5.7 shows the probability $p_D^{(\delta)}$ of falsely detecting an incorrect stimulus within one single oscillation cycle as a function of the distance δ between the correct stimulus and the falsely detected stimulus for each empirical neuron (for details see Section 2.1.3).

These probabilities are almost identical for the pure rate (red) and the combined phase and rate analysis (green) for the given empirical parameter set. Interestingly for almost all empirical neurons the probability of falsely detecting a stimulus with higher distance from the correct

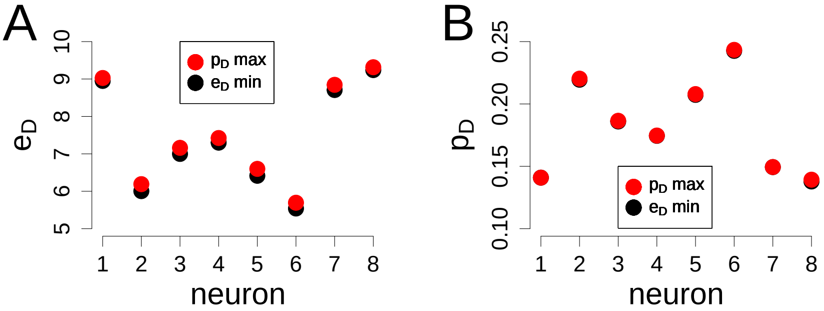


Figure 5.6: Comparison of the theoretically optimal neurons that maximize the detection probability (red) and the theoretically optimal neurons that minimize the detection error (black). A. Detection error. B. Detection probability.

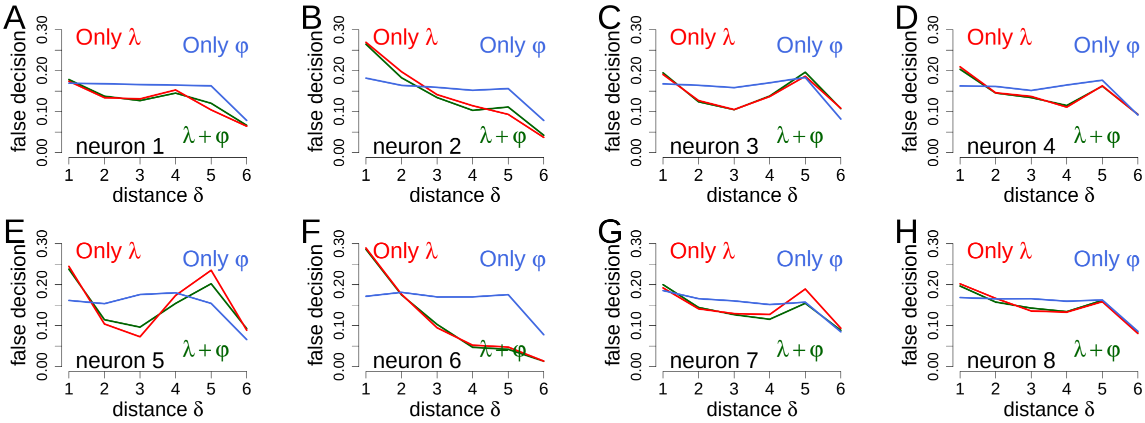


Figure 5.7: Average probability $p^{(\delta)}$ of falsely detecting a stimulus with the indicated distance δ after observing one cycle separately for each empirical neuron (10000 simulations per data point). Pure rate analysis shown in red, pure phase analysis in blue, bivariate analysis in green.

one decreases with the distance for a pure rate or a rate and phase analysis, especially for the interesting neuron 2 and 6. A pure phase analysis results in an almost flat curve, besides of distance six. However, considering the set of eights neurons simultaneously (Figure 5.8 B) the probability of falsely detecting a stimulus with higher distance from the correct one decreases rapidly with the distance (pure rate and bivariate analysis), implying that this set of eight neurons can already correctly identify the correct stimulus with a precision of about ±30° within one single oscillation cycle. A pure phase analysis results again in an almost flat curve.

5.3 Change point detection

In the following sections we analyze the performance of the empirical neurons in the change point detection. First we explain the procedure in a discrete change point setting, see Section

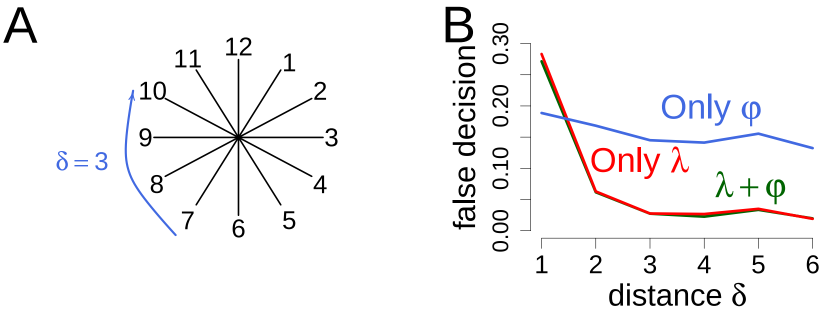


Figure 5.8: A. Illustration of the distance of the twelve measured stimuli. B. Average probability $p^{(\delta)}$ of falsely detecting a stimulus with the indicated distance δ after observing one cycle for the set of eight empirical neurons (10000 simulations per data point). Pure rate analysis shown in red, bivariate analysis in green.

5.3.1. Thereby we draw attention to the fact that a theoretical neuron which is optimal with respect to the detection probability is not necessarily optimal with respect to the change point detection. This especially concerns the optimal parameters of a pure rate code (Section 2.1.2.1). Therefore, we compare the empirical neurons only to a theoretic optimal neuron using rate and phase, and in case of a pure rate analysis, we consider the rate parameters of the optimal rate and phase neuron (which is not necessarily an optimal rate code).

In Section 5.3.2 we choose neuron 2 and compare its ability in the change point detection to a theoretic optimal rate and phase neuron. The improvement by the bivariate analysis compared to a pure rate analysis is more striking for the theoretic neuron and increases with higher firing rates.

As the set of eight neurons can already correctly identify the correct stimulus with a precision of about ±30° within one oscillation cycle, the detection ability of all neurons is very precise already for a pure rate analysis and the bivariate analysis improves the change point detection only in case of a small number of cycles or a short decision delay, cf. Section 5.3.3. Therefore, we consider the number of correct and of false detections dependent on the distance of stimuli and observe that imprecise phases can increase the probability of correctly detecting a change point especially in case of highly similar stimuli.

5.3.1 BOCD with a discrete uniform prior

In Section 5.3.2 we consider a discrete change point setting given by an empirical and a theoretical set of rate and phase parameters. Therefore, we need to extend the theoretical considerations of Section 4, where we used a normal prior distribution for the phase and a gamma prior distribution for the rate, cf. Section 4.3.1.3 and 4.3.2.1, to a discrete uniform prior distribution on the set of parameters. Here we describe how the predictive distribution can be derived required in the BOCD, see Equation (4.1) in Section 4.2.

In cycle k we observe the random vector $Z_k = (N_k, X^{(k)}_{1:N_k})$, where N_k denotes the number of spikes and $X^{(k)}_{1:N_k}$ the spike times. Again we use the BOCD to detect changes in the stimulus. To calculate the crucial run length distribution, we need to determine the predictive distribution. As prior distribution we assume now a discrete uniform distribution on the set

$\{(\lambda_1, \varphi_1), \cdots, (\lambda_S, \varphi_S)\}$, i.e.,

$$\Theta_0 := (\Lambda_0, \Phi_0) \sim \mathit{Unif}\left(\{(\lambda_1, \varphi_1), \cdots, (\lambda_S, \varphi_S)\}\right).$$

Let $\pi(\cdot)$ denote the prior distribution of Θ_0 and $\theta_s := (\lambda_s, \varphi_s)$, and suppose that we observe $k+1$ cycles without a change point. Then the predictive distribution is given by

$$p(Z_{k+1} = z_{k+1} \,|\, Z_{1:k} = z_{1:k}) = \sum_{s=1}^{S} p\left(Z_{k+1} = z_{k+1} \,|\, \Theta_0 = \theta_s\right) \mathbb{P}\left(\Theta_0 = \theta_s \,|\, Z_{1:k} = z_{1:k}\right).$$

The posterior distribution can be determined by

$$\begin{aligned}\mathbb{P}\left(\Theta_0 = \theta_s \,|\, Z_{1:k} = z_{1:k}\right) &= \frac{p(Z_{1:k} = z_{1:k} \,|\, \Theta_0 = \theta_s)\mathbb{P}(\Theta_0 = \theta_s)}{p(Z_{1:k} = z_{1:k})} \\ &= \frac{p(Z_{1:k} = z_{1:k} \,|\, \Theta_0 = \theta_s)}{\sum_{\tilde{s}=1}^{S} p(Z_{1:k} = z_{1:k} \,|\, \Theta_0 = \theta_{\tilde{s}})}.\end{aligned}$$

Since the BOCD is a recursive algorithm, the posterior distribution can also be determined recursively, i.e.,

$$\mathbb{P}\left(\Theta_0 = \theta_s \,|\, Z_{1:k} = z_{1:k}\right) = \frac{p(Z_k = z_k \,|\, \Theta_0 = \theta_s)\mathbb{P}\left(\Theta_0 = \theta_s \,|\, Z_{1:(k-1)} = z_{1:(k-1)}\right)}{p(Z_{1:(k-1)} = z_{1:(k-1)})},$$

where the first term of the numerator is given by the sampling model, the second term by the recursion and the denominator is just a normalization factor.

2 stimuli Here we apply the BOCD ($a_0 = b_0 = 1$) to the discrete change point setting of $S = 2$ stimuli. In Figure 5.9 A and B we consider $\boldsymbol{\lambda} = (1, 4)$, in Figure 5.9 C and D $\boldsymbol{\lambda} = (2, 4)$. In both cases we consider $\boldsymbol{\varphi} = (0, 0.75)$, whereas the order does not affect the change point detection.

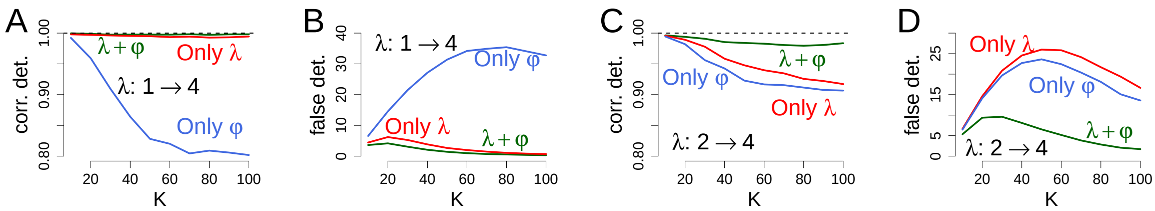


Figure 5.9: Discrete change point detection for 2 stimuli. Results of the BOCD considering spike trains of length $K = 10, \ldots, 100$ with one change point at $K/2$. In half of the simulations (1000 per data point) we consider a change of stimulus 1 to stimulus 2 and vice versa. A and B. $\boldsymbol{\lambda} = (1, 4)$ and $\boldsymbol{\varphi} = (0, 0.75)$. C and D. $\boldsymbol{\lambda} = (2, 4)$ and $\boldsymbol{\varphi} = (0, 0.75)$. A,C (B,D). Average number of correctly (falsely) detected change points as a function of the length K. Pure rate analysis shown in red, pure phase analysis in blue and bivariate analysis in green.

Basically the use of a discrete prior distribution on the rate and phase parameters give comparable results to the use of continuous prior distributions on rate and phase (Section

4.3.3): The bivariate analysis increases the number of correctly detected change points and decreases the number of falsely detected change points compared to a pure rate or pure phase analysis. However, the discrete setting reduces the number of false detections in case of a pure rate code (red) for a change $\lambda : 1 \to 4$ (Figure 5.9 B), but still does a lot of false detections for a change $\lambda : 2 \to 4$ (Figure 5.9 D). In general the number of correctly detected change points (Figure 5.9 A and C) is strikingly increased compared to the use of continuous prior distributions, as we now know the magnitude of possible changes. This is consistent to our results about mixtures of conjugate prior distributions (Section 4.4.2), which almost corresponds to a discrete prior distribution if we choose very precise prior distributions. But note there we need to adjust the prior parameters of the change point probability (b_0 large), as theoretical many small changes are possible.

Problems of rate coding In the change point setting we compare for $S = 3$ and $S = 7$ stimuli different parameter sets (optimal rate code, optimal rate and phase code) that maximize the detection probability, cf. Section 2.1.2. In Figure 5.10 A and B we consider $S = 3$ stimuli, with $\lambda_M = 4$ and $\varphi_M = 0.75$, and the optimal rate code (red dotted line)

$$\text{Opt } \lambda\text{:} \quad \boldsymbol{\lambda} = (1, 2, 4), \quad p_D = 0.6,$$

the optimal rate and phase code (green line)

$$\lambda + \varphi\text{:} \quad \boldsymbol{\lambda} = (\sqrt{2}, 4, 4), \quad \boldsymbol{\varphi} = (0, \varphi_M, 0), \quad p_D = 0.707,$$

and the pure rate analysis of the optimal rate and phase code (red line)

$$\text{Only } \lambda\text{:} \quad \boldsymbol{\lambda} = (\sqrt{2}, 4, 4), \quad p_D = 0.587.$$

For the parameters of the optimal rate and phase code the BOCD detects almost every change

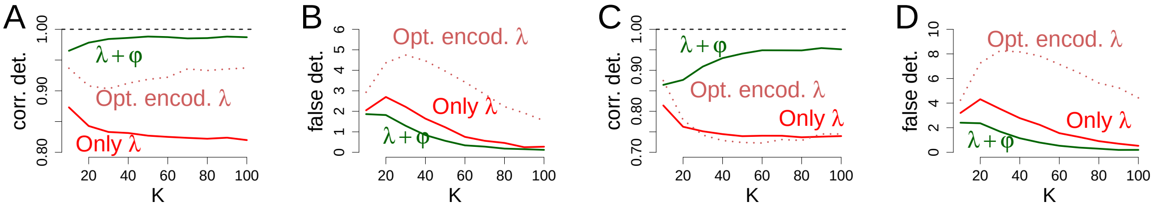


Figure 5.10: Discrete change point detection for $S = 3$ ($\lambda_M = 4$ and $\varphi_M = 0.75$) and S=7 ($\lambda_M = 7$ and $\varphi_M = 0.75$) stimuli. Results of the BOCD considering spike trains of length $K = 10, \ldots, 100$ with one change point at $K/2$. In each simulation (10000 per data point) we uniformly choose two stimuli s_1 and s_2 without replacement and consider a change of stimulus s_1 to stimulus s_2. A and B. $S = 3$ stimuli: Opt. λ: $\boldsymbol{\lambda} = (1, 2, 4)$; $\lambda + \varphi$: $\boldsymbol{\lambda} = (\sqrt{2}, 4, 4)$ and $\boldsymbol{\varphi} = (0, \varphi_M, 0)$; only λ: $\boldsymbol{\lambda} = (\sqrt{2}, 4, 4)$. C and D. $S = 7$ stimuli: Opt. λ: $\boldsymbol{\lambda} = (1, 2, \ldots, 7)$; $\lambda+\varphi$: $\boldsymbol{\lambda} = (1.44, 1.44, 3.48, 3.48, 7, 7, 7)$ and $\boldsymbol{\varphi} = (0, \varphi_M, 0, \varphi_M, 0, \varphi_M/2, \varphi_M)$; only λ: $\boldsymbol{\lambda} = (1.44, 1.44, 3.48, 3.48, 7, 7, 7)$. A,C (B,D). Average number of correctly (falsely) detected change points as a function of the length K. Pure rate analysis shown in red, pure phase analysis in blue and bivariate analysis in green.

point correctly and in case of at least $K \approx 60$ cycles does almost no false detections. Not surprisingly in case of the optimal rate code (dotted line) the BOCD detects more change points correctly compared to the pure rate analysis of the optimal rate and phase code (red line), but does a lot more false detections. More convincing for $S = 7$ stimuli and $\lambda_M = 7$ and $\varphi_M = 0.75$ (Figure 5.10 C and D). The optimal rate code (dotted line, cf. Section 2.1.2.1) is

$$\text{Opt } \lambda: \quad \boldsymbol{\lambda} = (1, 2, \ldots, 7), \quad p_D = 0.368,$$

the optimal rate and phase code (green line)

$$\lambda + \varphi: \quad \boldsymbol{\lambda} = (1.44, 1.44, 3.48, 3.48, 7, 7, 7), \quad \boldsymbol{\varphi} = (0, \varphi_M, 0, \varphi_M, 0, \varphi_M/2, \varphi_M), \quad p_D = 0.485,$$

and the pure rate analysis of the optimal rate and phase code (red line)

$$\text{Only } \lambda: \quad \boldsymbol{\lambda} = (1.44, 1.44, 3.48, 3.48, 7, 7, 7), \quad p_D = 0.353.$$

Here in case of the optimal rate code and in case of a pure rate analysis of the optimal rate and phase code, we observe almost the same number of correctly detected change points, but in case of the optimal rate code we have a much higher number of falsely detected change points. This is due to the similar rate parameters between neighbored stimuli. In case of an optimal rate and phase code a change between stimulus 3 and 4 can not be detected with a pure rate analysis, but therefore a change between stimulus 4 and 5 can be detected clearly. Instead, in case of an optimal rate code a change between stimulus 3 and 4 is possible to detect, but it is quite difficult with a moderate number of cycles and also a change between stimulus 4 and 5 is difficult to detect.

Therefore, in change point detection an optimal rate and phase code is more appropriate if we restrict to only rate, as we automatically have pairs of stimuli with the same rate parameters. Thus this emphasizes again the importance of imprecise phases, i.e., imprecise phases increase the detection probability while simultaneously enable an improved and more considerably robust change point detection.

5.3.2 One representative empirical neuron

Here we consider neuron 2 and compare its ability in the change point detection to a theoretic neuron that maximizes the detection probability (with the same $\lambda_M^{(2)}$ and $\varphi_M^{(2)}$ as the empirical neuron). Neuron 2 is chosen here as it shows a relatively large range of rate and phase parameters across stimuli, providing the possibility of stimulus encoding by rate and phase parameters. Among the eight reported neurons, neuron 5 and 6 showed similar parameter structures and yielded comparable results (data not shown). The remaining five neurons showed too small firing rates to be suitable for consideration in a single neuron context, cf. Section 5.2.1.

For convenience, we restrict the analysis to those eight out of 12 stimuli to which this neuron showed firing rates of at least 0.5 spikes per oscillation cycle. The reduction to eight stimuli also allows a comparison of this empirical neuron with the optimal combination of rate and phase parameters in the given parameter range derived according to Section 2.1.2.3 (see Figure 5.11 B).

Because of the high number of stimuli in combination with relatively small numbers of spikes and moderate phase differences, we focus on the comparison of the analysis based on rate alone with the bivariate analysis based on rate and phase, and do not consider the analysis based on phase alone.

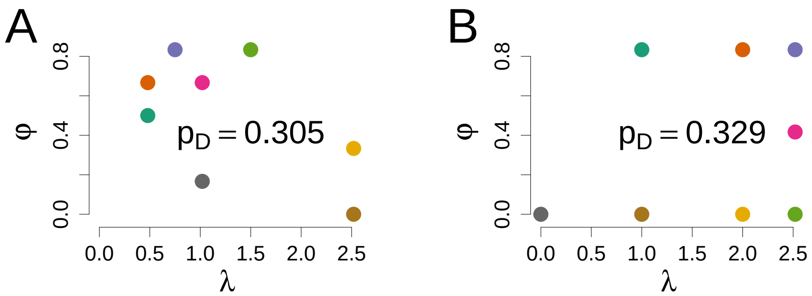


Figure 5.11: A. Observed parameters of empirical neuron 2 reported in (Havenith et al., 2011). B. The parameters optimizing detection probability for the given parameter range.

Detection probability If we plug in the phase and rate parameters of Figure 5.11 to derive the probability to detect the correct stimulus, the empirical neuron shows almost the same detection probability ($p_D = 0.305$) as the optimal parameter set ($p_D = 0.329$). This close-to-optimal behavior of the empirical neuron in the sense of rate and phase coding is particularly interesting considering that the theoretical neuron even uses the complete parameter range, including also a nullstimulus, which necessarily increases detection probability.

Furthermore, note that the rate parameters that maximize the probability of correct stimulus detection within the given parameter range when ignoring phase is identical to the rate parameters shown in Figure 5.11 B (see 'A note on the case of small rate and many stimuli when $\lambda_M \leq S$', page 26, and Figure 2.2 C). Based only on rate, the optimal detection probability is $p_D = 0.262$ (theoretical neuron) and $p_D = 0.251$ (empirical neuron). Hence the additional phase parameter increases the detection probability by similar amounts, i.e., 25.4% for the theoretical neuron and 21.5% for the empirical neuron.

Change point detection Here we compare the empirical and theoretical parameter sets in Figure 5.11 with respect to the performance in the change point detection task. Note that the parameter combination of the theoretical neuron is only optimal with respect to the detection probability and not necessarily with respect to change point detection. However, we consider this neuron a suitable candidate for comparison as pure change point detection will in practice be important only if it is accompanied by correct stimulus detection.

In order to apply the change point detection methods to the example data set of eight phase and rate parameters, we extend our techniques for application to a discrete set of stimuli. We assume now that at a change point a new stimulus and its underlying rate and phase parameters are chosen uniformly at random from the set of eight parameter combinations. In the algorithm, we then use a discrete uniform prior distribution on this set, and derive the predictive distribution accordingly (for details see Section 5.3.1).

Figures 5.12 A and B show the results of the BOCD (circles) and of the BOCD with online decision (curves) as a function of the decision delay for the empirical neuron (solid) and for the theoretical neuron (dashed). We simulated spike trains of length $K = 100$ with exactly one change point at $K/2$. In each of 10000 simulations two stimuli were drawn randomly from the set of eight stimuli, where the first and second part of the spike train corresponded to the parameters of the first and second stimulus, respectively.

The bivariate analysis based on rate and phase (green) showed a higher number of correctly detected change points than the analysis based on rate alone (red, panel A), while the number of falsely detected change points was reduced as compared to the pure rate analysis (panel B). As these parameter sets showed a relatively high number of falsely detected change points due to the small spike numbers and the high number of stimuli, we also performed analogous simulations in which we multiplied all rate parameters by four (panels C and D) to illustrate the effect of a number of neurons with similar response characteristic. In that case, the improvement in the number of correctly detected change points was even stronger (panel C), while also the number of falsely detected change points decreased (panel D). Almost no change points were falsely detected for decision delays of at least five.
Thus, the bivariate analysis using rate and phase parameters could increase the number of correctly detected change points as well as increase robustness by decreasing the number of falsely detected change points in the stimulus.

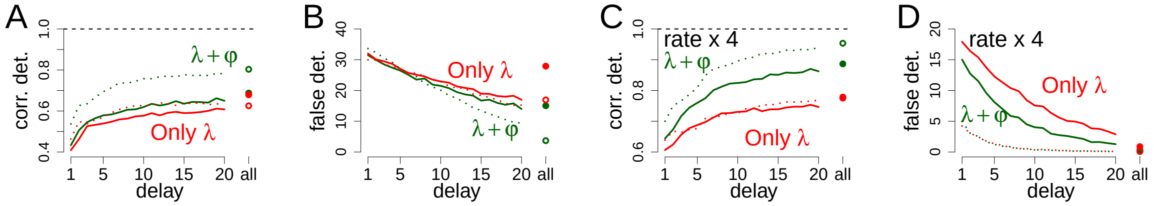


Figure 5.12: Fraction of correctly detected change points (A,C) and number of falsely detected change points (B,D) resulting from application of the BOCD (circles) and the BOCD with online decision (curves) to simulated sequences of length $K = 100$ with exactly one change point at $K/2 = 50$ (10000 simulations per data point). X-axis indicates duration of decision delay for BOCD with online decision. Pure rate analysis shown in red, bivariate analysis in green. In each simulation, the parameters of two stimuli before and after the change point were drawn randomly out of the set of all considered eight stimuli, using in A and B the empirical (solid lines) and the theoretical (dashed lines) neuronal parameter combinations in Figure 5.11. C and D. Analogous simulations, where the rate parameters of the empirical and theoretical neuron were multiplied by 4.

5.3.3 Change point detection dependent on distance

Here we consider the set of eight neurons simultaneously and analyze its ability in the discrete change point detection (all $S = 12$ stimuli). In Section 5.2.2 we have already seen that all neurons simultaneously can distinguish stimuli with a distance of $\delta > 1$ quite safe in only one cycle. Therefore, it is not surprising that a change between two randomly chosen stimuli ($11 \cdot 12 = 132$ possible stimuli combinations and only $2 \cdot 12 = 24$ with a distance of $\delta = 1$) can be detected almost ever, even for a small number of cycles (Figure 5.13 A) or online with a small delay (Figure 5.13 C) and with only few falsely detected change points (Figure 5.13 C and D). Therefore we further investigated the detection of change points in the spike train caused by changes in the stimulus as a function of the distance between the stimuli before and after the change. To that end we applied the BOCD with online decision with a fixed decision delay of $d = 5$ to spike trains of length $K = 100$ cycles with exactly one change point

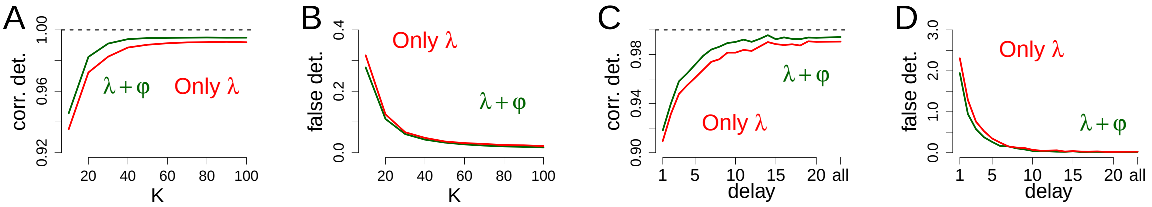


Figure 5.13: Discrete change detection using the set of eight empirical neurons. A change between two randomly chosen stimuli occurs at $K/2$ (10000 simulations per data point). A and B. Results of the BOCD considering spike trains of length $K = 10, \ldots, 100$. C and D. Results of the BOCD with online decision considering spike trains of length $K = 100$. A and C (B and D). Average number of correctly (falsely) detected change points.

in the middle of the spike train. Again, change points between stimuli with a distance δ of at least three (corresponding to 90°) could be detected almost with probability 1. However, stimuli with smaller distances showed considerably smaller probabilities of correct change point detection (Figure 5.14 A). This applied both for the pure rate (red) and the combined rate and phase analysis (green). However, the bivariate analysis could increase the probability of change point detection particularly for small distances, and it reduced the probability of falsely detecting a change point and thus increased robustness for all stimulus distances (panel B).

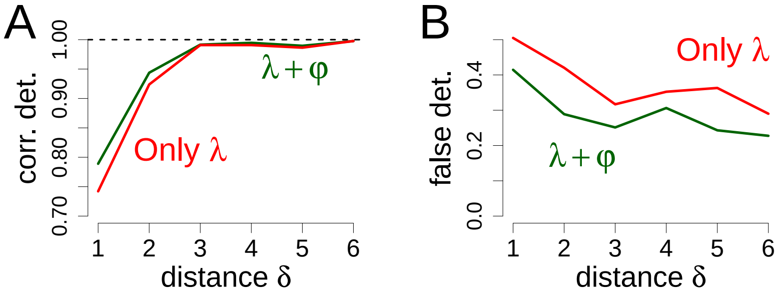


Figure 5.14: Change point detection as a function of stimulus distance. Results of BOCD with online decision and a fixed decision delay of $d = 5$, considering spike trains of length $K = 100$ with one change point at $K/2 = 50$. A. (B.). Average number of correctly (falsely) detected change points as a function of the distance between the stimuli before and after the change (10000 simulations per data point). Pure rate analysis shown in red, bivariate analysis in green.

Taken together, the results of Section 5 suggest that near-optimal parameter combinations of rate and phase do exist in the brain, and that therefore, the contribution of imprecise phases to information processing as investigated theoretically in Sections 2.1.2 and 4.3.3 can also be observed empirically. Particularly, imprecise phases can increase the probability of correctly detecting a change point especially in case of high firing rates or highly similar stimuli, and they can reduce the probability of falsely detecting a change point.

Chapter 6

Conclusion

Precise phase of spiking can carry sensory information beyond the information contained in the spike count (Srivastava et al., 2017; Kayser et al., 2009; Nemenman et al., 2008; Thorpe et al., 2001). This is theoretically clear if an accurate measurement of the spike position is possible. However, neuronal firing often exhibits a high degree of variability, or noise, yielding mean phases that can be measured in the long run but not in short time scales, such as in individual oscillation cycles (Havenith et al., 2011; Bizley et al., 2010; Lorenzo et al., 2009; Nelken et al., 2005). It is unclear to which degree noisy, or imprecise phases may be important for neuronal information processing in addition to or as compared to the signal component of firing rate. This question becomes particularly important considering the high speed of neuronal information processing, which is assumed to be based on only a few milliseconds, or oscillation cycles within each processing step (Osram et al., 1999; Gautrais and Thorpe, 1998; Abeles, 1994).

We have used a parsimonious stochastic spiking model, which in a single oscillation cycle is reduced to only two parameters corresponding to the signal components of rate λ and phase φ. Thereby the number of spikes is assumed Poisson distributed with parameter λ, while the position of each spike is placed independently according to a normal distribution with mean φ and unit variance.

The present approach based on the simple stochastic model has a number of advantages. First, the model contains exactly two signal parameters describing directly the rate and the phase, and it describes the properties of individual oscillation cycles. This allows the investigation of two quantities, the probability of correct stimulus detection, and the probability of correct change point detection, as a function of these signal parameters and within short periods of time such as individual oscillation cycles. Second, this allows optimization of the signal parameters with respect to these quantities and comparison of pure rate, pure phase and combined codes. Third, parameter estimation is simple and straightforward, where spiking patterns with similar number and phase of spikes are automatically assigned similar estimators. The procedure also works without artificial introduction of a temporal binning structure which might affect the results.

Within this model, we have investigated optimal combinations of rate and/or phase parameters that maximize the probability of correct stimulus detection, p_D (Section 2), more in detail for the case of a single neuron and a single oscillation cycle (Section 2.1). Depending on the parameter range of rate and phase parameters, the resulting optimal parameter combinations comprised pure rate codes in cases with highly imprecise phases and high rate differences, pure

phase codes in cases with highly precise phases and moderate or large rates, and combined codes. In general, the increase in p_D when adding imprecise phase coding to pure rate coding increases with the number of stimuli. In a case of eight stimuli, this increase ranged up to 30% in the neurophysiologically plausible parameter ranges considered here.
Note that we therefore considered a single oscillation cycle. If the information processing allows enough time to use for example two oscillation cycles, the optimal parameter combinations change, but they are roughly based on single cycles with twice as high rate parameters (Section 2.1.5). Due to the additional uncertainty of the spike allocation to the correct oscillation cycle, the increase in p_D when adding imprecise phase coding to pure rate coding decreases compared to a single cycle.
Another restraint of the present methods is that they focus on the coding structure of a single neuron for a relatively large number of stimuli. This is because the theoretically optimal parameter combination needs to be derived numerically and comprises as many as $2S$ parameters for a single neuron. In this respect, deriving the optimal parameter combination for only two neurons is already a numerically difficult task, where the computational cost is heavily increasing in the number of stimuli and the number of neurons (Section 2.2). Basic considerations however suggest that coding can be optimized easily when combining several neurons. For example, for two neurons and two stimuli, each neuron can increase its rate for another stimulus, which results in a highly robust coding based on rate alone. However, in cases of high numbers of stimuli ($S \geq 2^M$), the phase parameter still show increasing relevance.

In addition, we found that imprecise phases can improve the process of detecting changes in the stimulus (Section 4). In particular, including the phase parameter in the change point analysis in addition to the rate parameter increases the probability of correct change point detection. More importantly, it considerably decreases the probability of false alarms, thereby massively increasing robustness of change point detection. This holds for both, offline and fast online decision processes with only a short decision delay investigated in Section 4.3. To obtain a robust change point detection in the phase it is crucial to incorporate changes in the spike precision if they occur (Section 4.5). But even with an unknown and random spike precision, imprecise phases can improve the change point detection. Furthermore, change point detection based on pure rate or pure phase analysis can perform very similar in case of highly precise prior information (Section 4.4)
Note that we assume that all spikes can be assigned to the correct oscillation cycle, and that the temporal delay with respect to a theoretical onset of this cycle is known or can be measured precisely. Considering noisy processes in empirical recordings, these assumptions must be considered artificial, but they were used in order to investigate theoretical optimality. In addition, information about the theoretical onset of a cycle may be unnecessary in practice in the presence of multiple active neurons. In this case, only the delay of spikes of different neurons will be of practical relevance, which may be more easily tractable in a neurophysiological way.

Our theoretical results suggest that including imprecise phases can not increase the detection probability for many neurons and relative small number of stimuli. The importance of imprecise phases in such a setting is explored in Section 5, where we have applied our methods to parameters extracted from empirical spike train recordings of eight neurons with respect to 12 stimuli. The results suggest that near-optimal combinations of rate and phase parameters may be implemented in the brain, and that phase parameters can particularly increase the

sensitivity and robustness of change point detection in cases of highly similar stimuli.
In summary, the simple stochastic model with a rate and a phase parameter suggests that the use of imprecise spike timing can not only increase the probability of correct stimulus detection, but also increase the number of correctly detected changes in the stimulus. More importantly, adding a phase parameter can increase robustness, i.e., decrease the number of false alarms in the detection of changes in the signal. In addition, the model allows the investigation of basic coding principles on the level of empirical recordings. In the empirical parameters extracted from Havenith et al. (2011) for example, stimuli of sufficiently large difference could be correctly distinguished almost with probability one within only a single oscillation cycle - even only on the basis of rate parameters. In this setting, changes between highly similar stimuli could be detected more reliably by the additional consideration of phase parameters. These results suggest that small and imprecise phases can contribute to information processing, increasing the probability and precision of correct stimulus detection as well as enabling robust detection of changes in the input signal.

Additional information

A Parameter range in data

Here we explain the choice of the parameter range for the rate and phase parameters throughout the thesis, which is $\lambda \in [0, 4]$ and $\varphi \in [0, 0.75]$ (cmp. Sections 2 and 4). We focus on the parameters in the specific context of small imprecise phases, reported in (Havenith et al., 2011; Schneider, 2008; Schneider and Nikolić, 2006).

Concerning the rate parameter, (Havenith et al., 2011) report single unit activity with a mean rate of 18 ± 15 Hz and an average length of an oscillation cycle of about 60 ms, resulting in an approximate maximal rate parameter of about 3 spikes per cycle or slightly more (see also Figure 5C in (Havenith et al., 2011)). The firing rates reported in (Schneider, 2008) (Figure 3) ranged up to 3.7 spikes per cycle.

Phase parameters in the used range of about $\varphi \in [0, 0.75]$ have been extracted previously by fitting a similar stochastic model reported in (Schneider, 2008) to neuronal firing activity recorded in parallel in cat primary visual cortex under visual stimulation (Schneider and Nikolić, 2006). Instead of the normal distribution of spike times, an exponential distribution with temporal parameter τ was used, which was estimated in the range of $3.5 - 8$ ms. If we consider $Y_1 \sim \exp(\tau_{\min})$ and $Y_2 \sim \exp(\tau_M)$ the spike times corresponding to the maximal phase difference using exponential distributions and analogously $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim \mathcal{N}(\varphi_M, 1)$ for normal distributions, we observe

$$\begin{aligned} \mathbb{E}[Y_2 - Y_1] &= \tau_M - \tau_{\min} \quad \text{and} \quad \mathbb{E}[X_2 - X_1] = \varphi_M - 0, \\ \mathbb{V}ar[Y_2 - Y_1] &= \tau_M^2 + \tau_{\min}^2 \quad \text{and} \quad \mathbb{V}ar[X_2 - X_1] = 2. \end{aligned}$$

We then approximate the maximal standardized phase by

$$\varphi_M = \frac{\mathbb{E}[X_2 - X_1]}{\sqrt{\mathbb{V}ar[X_2 - X_1]/2}} \approx \frac{\mathbb{E}[Y_2 - Y_1]}{\sqrt{\mathbb{V}ar[Y_2 - Y_1]/2}} = \frac{\tau_M - \tau_{\min}}{\sqrt{(\tau_M^2 + \tau_{\min}^2)/2}} \approx 0.73.$$

A similar result is obtained by setting the detection probability, cf. Section 2.1.1, equal for the two models in case of two stimuli and identical rates. In this case, in both models we have two decision areas, where we decide for stimulus 1, if we observe a spike time less than the optimal decision bound c. In case of normally distributed spike times the optimal decision bound is $c = \varphi_M/2$. In case of exponentially distributed spike times the optimal decision bound c can be determined by solving $\tau_M e^{-\tau_M c} = \tau_{\min} e^{-\tau_{\min} c}$, i.e.,

$$c = \frac{\log(\tau_M/\tau_{\min})}{1/\tau_{\min} - 1/\tau_M}.$$

With that choice equating the detection probabilities for both models, i.e.,

$$\mathbb{P}(Y_1 \le c) + \mathbb{P}(Y_2 > c) = \mathbb{P}(X_1 \le \varphi_M/2) + \mathbb{P}(X_2 > \varphi_M/2),$$

yields a maximal phase difference $\varphi_M \approx 0.76$.
Similarly, the phases of individual neurons reported in (Havenith et al., 2011, Figure 5B) show ranges of about 6 ms, where the standard deviation σ is estimated roughly as $\hat{\sigma} = 6$ from Figure S4 in (Havenith et al., 2011), yielding an approximate maximal range of about $\varphi_M = 1$ or slightly less.
Note that these choices clearly refer to the respective experimental context, where however generalized results concerning smaller or larger phases can also be found in Section 2.1.2.

B Basic definitions and properties

Here we summarize some basic definitions and properties of distributions, which we draw on in the thesis. We do not check that the probability density or mass functions are well-defined and integrate to one. Further we restrain on the less commonly used distributions, the notations of all relevant distributions can be found in the list of abbreviations (page 197).

Definition B.1. *A continuous random variable X is said to have a Beta distribution with shape parameters $a, b > 0$, denoted as $X \sim Beta(a, b)$, if its probability density function is given by*

$$f(x) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, & \textit{if } x \in [0,1], \\ 0, & \textit{otherwise,} \end{cases}$$

where $\Gamma(z) := \int_0^\infty y^{z-1}e^{-y}dy$, $z > 0$, is the gamma function.

Remark B.2. *Let $z > 0$, then it holds for the gamma function*

$$\Gamma(z+1) = z\Gamma(z).$$

Proof.

$$\begin{aligned} \Gamma(z+1) = \int_0^\infty y^z e^{-y} dy &= \left[-y^z e^{-y}\right]_0^\infty + \int_0^\infty z y^{z-1} e^{-y} dy \\ &= \lim_{y\to\infty} \left(-y^z e^{-y}\right) + z\int_0^\infty y^{z-1}e^{-y}dy \\ &= z\Gamma(z). \end{aligned}$$

□

As $\Gamma(1) = 1$, we know with Remark B.2 that $\Gamma(k+1) = k!$ for $k \in \mathbb{N}$.

Remark B.3. *The expected value of $X \sim Beta(a, b)$ is*

$$\mathbb{E}[X] = \frac{a}{a+b}$$

Proof. Applying Remark B.2 it can be derived as follows

$$\begin{aligned}\mathbb{E}[X] &= \int_0^1 x\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}dx \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 x^{a}(1-x)^{b-1}dx \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a+b+1)}\frac{\Gamma(a+1)}{\Gamma(a)} = \frac{a}{a+b}.\end{aligned}$$

□

Definition B.4. *A continuous random variable X is said to have a Gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$, denoted as $X \sim Gamma(\alpha, \beta)$, if its probability density function is given by*

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, & \textit{if } x > 0, \\ 0, & \textit{otherwise.} \end{cases}$$

Remark B.5. *The expected value of $X \sim Gamma(\alpha, \beta)$ is*

$$\mathbb{E}[X] = \frac{\alpha}{\beta}$$

and the variance is

$$\mathbb{V}ar[X] = \frac{\alpha}{\beta^2}.$$

Proof. First we calculate the k-th moment

$$\begin{aligned}\mathbb{E}[X^k] &= \int_0^\infty x^k\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}dx = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha+k-1}e^{-\beta x}dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+k)}{\beta^{\alpha+k}} = \frac{(\alpha+k-1)(\alpha+k-2)\cdots\alpha}{\beta^k}.\end{aligned}$$

With that we directly obtain the mean. The variance is

$$\mathbb{V}ar[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{\alpha(\alpha+1)}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}.$$

□

Remark B.6. *If we let $\alpha = 1$ in Definition B.4, we obtain*

$$f(x) = \begin{cases} \beta e^{-\beta x}, & \textit{if } x > 0, \\ 0, & \textit{otherwise,} \end{cases}$$

which is an Exponential distribution with rate β and denoted as $Exp(\beta)$.

Remark B.7. *Let $E_1, \dots, E_k \sim Exp(\beta)$ and independent. Then*

$$\sum_{i=1}^{k} E_i \sim Gamma(k, \beta).$$

Proof. This can be easily seen by induction. The initial step is noticed in Remark B.6. Consider $Y := \sum_{i=1}^{k-1} E_i$ and $Z := Y + E_k$. According to the induction hypothesis $Y \sim Gamma(k-1, \beta)$. The probability density function of Z is

$$\begin{aligned} f_Z(z) &= \int_0^z f_Y(y) f_{E_k}(z-y) dy = \int_0^z \frac{\beta^\alpha}{\Gamma(k-1)} y^{k-2} e^{-\beta y} e^{-\beta(z-y)} dy \\ &= \frac{\beta^\alpha}{\Gamma(k-1)} e^{-\beta z} \int_0^z y^{k-2} dy = \frac{\beta^\alpha}{\Gamma(k-1)} e^{-\beta z} \frac{z^{k-1}}{k-1} = \frac{\beta^\alpha}{\Gamma(k)} z^{k-1} e^{-\beta z}. \end{aligned}$$

□

Remark B.8. *Let $X_1 \sim Gamma(p_1, m)$ and $X_2 \sim Gamma(p_2, m)$, $p_1, p_2, m > 0$ and independent. Then*

$$\frac{X_1}{X_1 + X_2} \sim Beta(p_1, p_2).$$

Proof. First note that the joint probability density function f of (X_1, X_2) is given by

$$f(x_1, x_2) = \frac{m^{p_1+p_2}}{\Gamma(p_1)\Gamma(p_2)} x_1^{p_1-1} x_2^{p_2-1} e^{-m(x_1+x_2)}, \quad x_1, x_2 \in (0, \infty).$$

Now we do the transformation $U := \frac{X_1}{X_1+X_2}$ and $V := X_1 + X_2$ with inverse $g^{-1}(U, V) = (UV, V(1-U))$. As the Jacobi matrix is given by

$$\mathcal{J} g^{-1}(u, v) = \begin{pmatrix} v & u \\ -v & 1-u \end{pmatrix}$$

the determinant of the Jacobi matrix is

$$det\ \mathcal{J} g^{-1}(u, v) = v(1-u) + uv = v.$$

Hence by the change of variables formula the probability density function h of the transformation is given by

$$\begin{aligned} h(u, v) &= det\ \mathcal{J} g^{-1} f\left(g^{-1}(u, v)\right) = v f(uv, v(1-u)) \\ &= v \frac{m^{p_1+p_2}}{\Gamma(p_1)\Gamma(p_2)} (uv)^{p_1-1} (v(1-u))^{p_2-1} e^{-m(uv+v(1-u))} \\ &= \frac{m^{p_1+p_2}}{\Gamma(p_1+p_2)} v^{p_1+p_2-1} \frac{\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)} u^{p_1-1} (1-u)^{p_2-1}, \quad u \in (0, \infty),\ v \in (0, 1), \end{aligned}$$

where the factor in v is the density function of a $Gamma(p_1 + p_2, m)$-distribution and the factor in u is the density of a $Beta(p_1, p_2)$-distribution. Moreover, $X_1 + X_2$ and $\frac{X_1}{X_1+X_2}$ are independent. □

Definition B.9. *A discrete random variable X is said to have a negative binomial distribution with parameters $k \in \mathbb{R}$ and $\eta \in (0,1)$, denoted as $X \sim \mathcal{NB}(k,\eta)$, if its probability mass function is given by*

$$\mathbb{P}(X = r) = \frac{\Gamma(k+r)}{r!\Gamma(k)}\eta^k(1-\eta)^r, \quad r \in \mathbb{N}.$$

Remark B.10. *For $X \sim \mathcal{NB}(k,\eta)$ and $k \in \mathbb{N}$ we obtain*

$$\frac{\Gamma(k+r)}{r!\Gamma(k)} = \frac{(k+r-1)!}{r!(k-1)!} = \binom{k+r-1}{r}$$

with the interpretation, that X is counting the number of failures r in a coin toss experiment with success probability η and k successes.

Definition B.11. *A continuous random variable X is said to have a Inverse-Gamma distribution with shape parameter $\gamma > 0$ and scale parameter $\delta > 0$, denoted as $X \sim \mathcal{IG}(\gamma,\delta)$, if its probability density function is given by*

$$f(x) = \begin{cases} \frac{\delta^\gamma}{\Gamma(\delta)} x^{\gamma-1} e^{-\delta/x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Remark B.12. *The expected value of $X \sim \mathcal{IG}(\gamma,\delta)$ is*

$$\mathbb{E}[X] = \frac{\delta}{\gamma-1}.$$

Proof.

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty x \frac{\delta^\gamma}{\Gamma(\delta)} x^{\gamma-1} e^{-\delta/x} dx \\ &= \frac{\delta}{\gamma-1} \int_0^\infty \frac{\delta^{\gamma-1}}{\Gamma(\delta-1)} x^{(\gamma-1)-1} e^{-\delta/x} dx = \frac{\delta}{\gamma-1}. \end{aligned}$$

□

Definition B.13. *A continuous random variable X is said to have a Student's t-distribution with location parameter $\mu \in \mathbb{R}$, scale parameter $\psi > 0$ and $\nu > 0$ number of degrees of freedom, denoted as $X \sim \mathcal{T}_\nu(\mu,\psi^2)$, if its probability density function is given by*

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\psi} \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\psi}\right)^2\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}.$$

Remark B.14. *Let $X \sim \mathcal{T}_\nu(\mu,\psi^2)$, than $\frac{X-\mu}{\psi} \sim \mathcal{T}_\nu(0,1)$, which is called standardized student's t-distribution.*

Remark B.15. *The variance of $X \sim \mathcal{T}_\nu(0,1)$, $\nu > 2$, is*

$$\mathbb{V}ar[X] = \frac{\nu}{\nu-2}.$$

Proof. Since the probability density function is symmetric to zero, we have $\mathbb{V}ar[X] = \mathbb{E}[X^2]$ and calculate with integration by parts and integration by substitution ($c_\nu := \Gamma\left(\frac{\nu+1}{2}\right)/(\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu})$)

$$
\begin{aligned}
\mathbb{E}[X^2] &= c_\nu \int_{-\infty}^{\infty} x^2 \left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx \\
&= c_\nu \left[x\left(-\frac{\nu}{\nu-1}\right)\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu-1}{2}} \right]_{-\infty}^{\infty} + c_\nu \int_{-\infty}^{\infty} \frac{\nu}{\nu-1}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu-1}{2}} dx \\
&= 0 + c_\nu \int_{-\infty}^{\infty} \frac{\nu}{\nu-1}\frac{\sqrt{\nu}}{\sqrt{\nu-2}}\left(1+\frac{t^2}{\nu-2}\right)^{-\frac{(\nu-2)+1}{2}} dt \\
&= \frac{\frac{\nu-1}{2}\Gamma\left(\frac{\nu-1}{2}\right)}{\frac{\nu-2}{2}\Gamma\left(\frac{\nu-2}{2}\right)\sqrt{\pi\nu}} \int_{-\infty}^{\infty} \frac{\nu}{\nu-1}\frac{\sqrt{\nu}}{\sqrt{\nu-2}}\left(1+\frac{t^2}{\nu-2}\right)^{-\frac{(\nu-2)+1}{2}} dt \\
&= \frac{\frac{\nu-1}{2}}{\frac{\nu-2}{2}}\frac{\nu\sqrt{\nu}}{(\nu-1)\sqrt{\nu}} \cdot \frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu-2}{2}\right)\sqrt{\pi(\nu-2)}} \int_{-\infty}^{\infty} \left(1+\frac{t^2}{\nu-2}\right)^{-\frac{(\nu-2)+1}{2}} dt \\
&= \frac{\nu}{\nu-2}\cdot 1 = \frac{\nu}{\nu-2}.
\end{aligned}
$$

□

C R-Codes

Here a collection of the most important R-Codes can be found to allow a reproduction of the results in this thesis. Our results about the optimal parameter in the stimulus encoding (Section 2) are mainly based on determination of the detection probability, see Function 1 for one neuron and Function 2 for M neurons. The exploration in the change point detection (Section 4) is realized with the BOCD (Section 4.2) and BOCD with online decision (Section 4.2.4), the implementation for the bivariate analysis can be found in Function 3 and Function 4. In Section 4.4 we apply the BOCD in context of mixtures of conjugate prior distributions. The calculation of the posterior weights, which need to be transmitted to the BOCD, can be found in Function 5. In Section 5 we consider a discrete prior distribution on the stimulus specific rate and phase parameters, which were found in empirical neurons. The adjustment of the BOCD to discrete weights can be found in Function 6.

```
pd_1_neur<-function(lamvec,phivec,sigma,maxspike){
  # Input:
  #  lamvec...   Vector of rates.
  #  phivec...   Vector of phases.
  #  sigma...    Standard deviation of the spike times.
  #  maxspike... Maximum considered spike number for the summation.
  # Output:
  #  pd...       Calculated detection probability
  lam<-lamvec[which(lamvec>0)]
  phi<-phivec[which(lamvec>0)]
  if (min(lamvec)<0){
    return(0)
    break
  }
  if (min(lamvec)>=0){
    p<-rep(0,length(lam))
    for (i in 1:length(p)){
      l<-1
      u<-maxspike
      if (length(which(phi==phi[i]))>1){
        ur<-which(phi==phi[i])[which(lam[which(phi==phi[i])]>lam[i])]
        if (length(ur)>0){
          u<-max(0,floor(min((lam[i]-lam[ur])/log(lam[i]/lam[ur]))))
        }
        lr<-which(phi==phi[i])[which(lam[which(phi==phi[i])]<lam[i])]
        if (length(lr)>0){
          l<-max(1,ceiling(max((lam[i]-lam[lr])/log(lam[i]/lam[lr]))))
        }
      }
      if (u>=l){
        minmu<-which(phi>phi[i])
        maxmu<-which(phi<phi[i])
        for (j in l:u){
          if (length(minphi)>=1 & length(maxphi)>=1){
            p[i]<-p[i]+dpois(j,lam[i])*max(0,(pnorm(min((-lam[i]+lam[minphi]
                 +j*log(lam[i]/(lam[minphi])))/((sqrt(j)/sigma)*(phi[minphi]-phi[i]))
                 -(sqrt(j)/sigma)*(phi[i]-phi[minphi])/2),0,1)
                 -pnorm(max((-lam[i]+lam[maxphi]+j*log(lam[i]/(lam[maxphi])))
                        /(sqrt(j)/sigma*(mu[maxphi]-phi[i]))
                        -sqrt(j)/sigma*(phi[i]-phi[maxphi])/2),0,1)))
          }
          if (length(minphi)==0 & length(maxphi)>0){
            p[i]<-p[i]+dpois(j,lam[i])*(1
                  -pnorm(max((-lam[i]+lam[maxphi]+j*log(lam[i]/(lam[maxphi])))
                        /(sqrt(j)/sigma*(phi[maxphi]-phi[i]))
                        -sqrt(j)/sigma*(phi[i]-phi[maxphi])/2),0,1))
          }
          if (length(maxphi)==0 & length(minphi)>0){
            p[i]<-p[i]+dpois(j,lam[i])
            *(pnorm(min((-lam[i]+lam[minphi]+j*log(lam[i]/(lam[minphi])))
                        /(sqrt(j)/sigma*(phi[minphi]-phi[i]))
                        -sqrt(j)/sigma*(phi[i]-phi[minphi])/2),0,1))
          }
          if (length(maxphi)==0 & length(minphi)==0){
            p[i]<-p[i]+dpois(j,lam[i])
          }
        }
        p[i]<-p[i]/length(which(abs(lam-lam[i])<10^(-5) & abs(phi-phi[i])<10^(-5)))
      }
    }
    pd<-1/(length(lam)+1+length(which(lamvec==0)))*(sum(p)+1)
    return(pd)
  }
}
```

Listing 1: Function to determine the detection probability for $M = 1$ neuron and and arbitrary number S of stimuli, cf. Lemma 2.1.3. The nullstimulus is already included.

```
pd_M_neur_simu<-function(lammatr,phimatr,simula){
  # Input:
  #  lammatr... Matrix of rates: row i represents the rate profil of neuron i,
  #                              column j the rate profil of stimulus j.
  #  phimatr... Matrix of phases.
  #  simula...  Number of simulations.
  # Output:
  #  pd...      Simulated detection probability
  pd<-0
  for (i in 1:length(lammatr[1,])){
    n<-list()
    Z<-list()
    for (k in 1:length(lammatr[,1])){
      n[[k]]<-rpois(simula,lammatr[k,i])
      Z[[k]]<-rnorm(simula,0,1)
    }
    for (j in 1:simula){
      left<-1
      right<-1
      for (k in 1:length(lammatr[,1])){
        left<-left*lammatr[k,i]^(n[[k]][j])*exp(-lammatr[k,i])*
              exp(-(phimatr[k,-i]-phimatr[k,i])*
              (sqrt(n[[k]][j])*Z[[k]][j])+n[[k]][j]*(phimatr[k,i]-phimatr[k,-i])^2/2)
        right<-right*(lammatr[k,-i])^(n[[k]][j])*exp(-lammatr[k,-i])
      }
      error<-left-right
      if (min(error)>=0){
        same<-length(which(error==0))
        pd<-pd+1/(same+1)
      }
    }
  }
  return((pd)/((length(lammatr[1,]))*simula))
}
```

Listing 2: Function to determine the detection probability for an arbitrary number M of neurons and S stimuli by simulations, cf. Lemma 2.2.2.

```
BOCD_rate_plus_phase<-function(X,mu0,tau0,alpha,beta,a0,b0,sigma){
  # Input:
  #  X...       List of spike times: Each element of the list is a vector.
  #                                  Element k corresponds to spike times of cycle k.
  #  mu0...     Expectation of the normal prior distribution of the phase.
  #  tau0...    Standard deviation of the normal prior distribution of the phase.
  #  alpha...   Shape parameter of the gamma prior distribution of the rate.
  #  beta...    Rate parameter of the gamma prior distribution of the rate.
  #  a0...      First shape parameter of the beta prior distr. for cp-probability.
  #  b0...      Second shape parameter of the beta prior distr. for cp-probability.
  #  sigma...   Standard deviation of the spike times.
  # Output:
  #  cp_pos...  Detected changepoint positions.
  T<-length(X)
  spikes<-unlist(lapply(X,length))
  runandX<-list() # P(r_t=i | X_(1:t))
  runandX[[1]]<-1
  rundandXandat<-list() # P(r_t=i , a_t=j | X_(1:t))
  rundandXandat[[1]]<-matrix(1,1,1)
  mu_t_j<-list()  # update mu
  mu_t_j[[1]]<-mu0
  tau_t_j<-list()  # update tau
  tau_t_j[[1]]<-tau0^2
  alpha_t_j<-list() # update alpha
  alpha_t_j[[1]]<-alpha
  beta_t_j<-list() # update beta
  beta_t_j[[1]]<-beta
  for (t in 2:T){
    # Update of the posterior distribution
    if (spikes[t-1]>0){
      mu_t_j[[t]]<-c(mu0,(mu_t_j[[t-1]]/tau_t_j[[t-1]]+sum(X[[t-1]])/sigma^2)
                          /(1/tau_t_j[[t-1]]+spikes[t-1]/sigma^2))
      tau_t_j[[t]]<-c(tau0^2,1/(1/tau_t_j[[t-1]]+spikes[t-1]/sigma^2))
    }
    if (spikes[t-1]==0){
      mu_t_j[[t]]<-c(mu0,mu_t_j[[t-1]])
      tau_t_j[[t]]<-c(tau0^2,tau_t_j[[t-1]])
    }

    alpha_t_j[[t]]<-c(alpha,alpha_t_j[[t-1]]+spikes[t-1])
    beta_t_j[[t]]<-c(beta,beta_t_j[[t-1]]+1)

    # Calculate changepoint probability
    rtat<-matrix(0,nrow=t,ncol=t)
    if (spikes[t]>0){
      densnorm<-1
      M<-mu_0
      Ta<-tau_0^2
      for (i in 1:spikes[t]){
        densnorm<-densnorm*dnorm(X[[t]][i],mean =M,sd = sqrt(Ta+sigma^2))
        M<-(M/Ta+X[[t]][i]/sigma^2)/(1/Ta+1/sigma^2)
        Ta<-1/(1/Ta+1/sigma^2)
      }
      rtat[1,]<-c(0,colSums(rundandXandat[[t-1]])
                  *dnbinom(spikes[t], size=alpha, prob=1-1/(beta+1))
                  *densnorm*(seq(2,t,1)-1+a0-1)/(t+a0+b0-2))
    }
    if (spikes[t]==0){
      rtat[1,]<-c(0,colSums(rundandXandat[[t-1]])
                  *dnbinom(spikes[t], size=alpha, prob=1-1/(beta+1))
                  *(seq(2,t,1)-1+a0-1)/(t+a0+b0-2))
    }
    # Calculate growth probabilities
    if (spikes[t]>0){
      densnorm<-1
      M<- mu_t_j[[t]][-1]
      Ta<-tau_t_j[[t]][-1]
      for (i in 1:spikes[t]){
        densnorm<-densnorm*dnorm(X[[t]][i],mean =M,sd = sqrt(Ta+sigma^2))
        M<-(M/Ta+X[[t]][i]/sigma^2)/(1/Ta+1/sigma^2)
        Ta<-1/(1/Ta+1/sigma^2)
      }
      densnorm<-densnorm*dnbinom(spikes[t], size=alpha_t_j[[t]][-1],
                                 prob=1-1/(beta_t_j[[t]][-1]+1))
    }
    if (spikes[t]==0){
      densnorm<-dnbinom(spikes[t], size=alpha_t_j[[t]][-1],
                        prob=1-1/(beta_t_j[[t]][-1]+1))
    }
    rtat[2:t,-t]<-rundandXandat[[t-1]]
                  *(matrix(t-seq(2,t,1)+b0,nrow=t-1,ncol=t-1,byrow=T))
                  /(t+a0+b0-2)*matrix(densnorm,nrow=t-1,ncol=t-1,byrow=F)

    # Scaling for a robust calculation
    rundandXandat[[t]]<-rtat/sum(rtat)
  }

  rundandX<-lapply(rundandXandat,function(x){rowSums(x)})
  rundandX<-lapply(rundandX,unlist)
  # Determine changepoints (take the most likely runlength at the last cycle)
  cp_pos<-c()
  rt<-max(which(rundandX[[T]]==max(rundandX[[T]])))
  cp_pos<-c(T-rt+1,cp_pos)
  t<-T-rt
  while (t>0){
    rt<-max(which(rundandX[[t]]==max(rundandX[[t]])))
    cp_pos<-c(t-rt+1,cp_pos)
    t<-t-rt
  }
  cp_pos<-cp_pos[-1]
  return(cp_pos)
}
```

Listing 3: BOCD (Adams and MacKay, 2007; Wilson et al., 2010) (Section 4.2), adjusted to our change point model, assuming a change in rate and phase simultaneously.

```
BOCD_online_decision_rate_plus_phase<-function(X,mu0,tau0,alpha,beta,a0,b0,sigma,delay
  ){
  # Input:
```

```
  #  X...       List of spike times: Each element of the list is a vector.
  #                                  Element k corresponds to spike times of cycle k.
  #  mu0...     Expectation of the normal prior distribution of the phase.
  #  tau0...    Standard deviation of the normal prior distribution of the phase.
  #  alpha...   Shape parameter of the gamma prior distribution of the rate.
  #  beta...    Rate parameter of the gamma prior distribution of the rate.
  #  a0...      First shape parameter of the beta prior distr. for cp-probability.
  #  b0...      Second shape parameter of the beta prior distr. for cp-probability.
  #  sigma...   Standard deviation of the spike times.
  #  delay...   Decision delay of the changepoint detection
  # Output:
  #  cp_pos...  Detected changepoint positions.
  cp_pos<-c()
  Mu_0<-mu0
  Tau_0<-tau0
  A0<-a0
  B0<-b0
  Alpha<-alpha
  Beta<-beta
  for (i in 1:(length(X)-1-delay)){
    z<-BOCD_rate_plus_phase(X[i:(i+delay+1)],Mu_0,Tau_0,Alpha,Beta,A0,B0,sigma)
    # No changepoint in the delay horizon
    if (length(z)==0){
      Mu_0<-(Mu_0/Tau_0^2+sum(X[[i]])/sigma^2)/(1/Tau_0^2+length(X[[i]])/sigma^2)
      Tau_0<-sqrt(1/(1/Tau_0^2+length(X[[i]])/sigma^2))
      Alpha<-Alpha+length(X[[i]])
      Beta<-Beta+1
      A0<-A0
      B0<-B0+1
    }
    # At least one changepoint in the delay horizon
    if(length(z)>0){
      # Change detection at the current position
      if (z[1]==2){
        cp_pos<-c(cp_pos,i+1)
        Mu_0<-mu0
        Tau_0<-tau0
        Alpha<-alpha
        Beta<-beta
        A0<-A0+1
      }
      # Changepoint not at the current position
      if (z[1]!=2){
        Mu_0<-(Mu_0/Tau_0^2+sum(X[[i]])/sigma^2)/(1/Tau_0^2+length(X[[i]])/sigma^2)
        Tau_0<-sqrt(1/(1/Tau_0^2+length(X[[i]])/sigma^2))
        Alpha<-Alpha+length(X[[i]])
        Beta<-Beta+1
        A0<-A0
        B0<-B0+1
      }
    }
  }
  return(cp_pos)
}
```

Listing 4: BOCD with online decision (Section 4.2.4), assuming a change in rate and phase simultaneously. This function access Function 3.

```
weights_stimuli_mixture_distribution<-function(X,mu0,tau0,alpha,beta,a0,b0,sigma){
  # Input:
  #  X...       List of spike times: Each element of the list is a vector.
  #                                  Element k corresponds to spike times of cycle k.
  #  mu0...     Vector of expectations of the normal prior of the phase mixture.
  #  tau0...    Vector of standard deviations of the normal prior of the phase mixture
     .
  #  alpha...   Vector of shape parameters of the gamma prior of the rate mixture.
#  beta...    Vector of rate parameters of the gamma prior of the rate mixture.
#  a0...      First shape parameter of the beta prior distr. for cp-probability.
#  b0...      Second shape parameter of the beta prior distr. for cp-probability.
#  sigma...   Standard deviation of the spike times.
# Output:
#  w_t_j...   Weights of the priors for each cycle k and possible run length.
#             (starting with an uniform prior)
T<-length(X)
k<-length(mu0) # number of stimuli
spikes<-unlist(lapply(X,length))
mu_t_j<-list()  # update mu
mu_t_j[[1]]<-mu0
tau_t_j<-list()  # update tau
tau_t_j[[1]]<-tau0^2
alpha_t_j<-list(list()) # update alpha
alpha_t_j[[1]]<-as.list(alpha)
beta_t_j<-list() # update beta
beta_t_j[[1]]<-as.list(beta)
w_t_j<-list(list()) # update weights
w_t_j[[1]]<-as.list(rep(1/k,k))
c_t_j<-list(list())
c_t_j[[1]]<-as.list(rep(1/k,k))
for (t in 2:T){
  # Update of the posterior weights
  mu_t_j[[t]]<-list()
  tau_t_j[[t]]<-list()
  alpha_t_j[[t]]<-list()
  beta_t_j[[t]]<-list()
  c_t_j[[t]]<-list()
  w_t_j[[t]]<-list()
  if (spikes[t-1]>0){
    for (j in 1:k){
      # Update phase
      mu_t_j[[t]][[j]]<-c(mu0[j],(mu_t_j[[t-1]][[j]]/tau_t_j[[t-1]][[j]]
                                  +sum(X[[t-1]])/sigma^2)/(1/tau_t_j[[t-1]][[j]]
                                                          +spikes[t-1]/sigma^2))
      tau_t_j[[t]][[j]]<-c(tau0[j],1/(1/tau_t_j[[t-1]][[j]]+spikes[t-1]/sigma^2))
      cj<-rep(0,length(mu_t_j[[t-1]][[j]]))
      for (i in 1:length(mu_t_j[[t-1]][[j]])){
        integrand <- function(x) {a<-c(); for(r in 1:length(x)) {
          a<-c(a,dnorm(x[r],mu_t_j[[t-1]][[j]][i],sqrt(tau_t_j[[t-1]][[j]][i]))
              *dmvnorm(X[[t-1]],rep(x[r],spikes[t-1]),diag(sigma,spikes[t-1])))}
            ;return(a)}
        cj[i]<-integrate(integrand, lower = -Inf, upper = Inf)[[1]]
      }
      c_t_j[[t]][[j]]<-cj
      # Update rate
      alpha_t_j[[t]][[j]]<-c(alpha[j],alpha_t_j[[t-1]][[j]]+spikes[t-1])
      beta_t_j[[t]][[j]]<-c(beta[j],beta_t_j[[t-1]][[j]]+1)
      cj<-rep(0,length(alpha_t_j[[t-1]][[j]]))
      for (i in 1:length(alpha_t_j[[t-1]][[j]])){
        integrand <- function(la) {dgamma(la,alpha_t_j[[t-1]][[j]][i],
                                beta_t_j[[t-1]][[j]][i])*dpois(spikes[t-1],la)}
        cj[i]<-integrate(integrand, lower = 0, upper = Inf)[[1]]
      }
      c_t_j[[t]][[j]]<-c_t_j[[t]][[j]]*cj
    }
  }
  # Determine posterior weights
  if (spikes[t-1]==0){
    for (j in 1:k){
      mu_t_j[[t]][[j]]<-c(mu_0[j],mu_t_j[[t-1]][[j]])
      tau_t_j[[t]][[j]]<-c(tau_0[j],tau_t_j[[t-1]][[j]])
      alpha_t_j[[t]][[j]]<-c(alpha[j],alpha_t_j[[t-1]][[j]]+spikes[t-1])
      beta_t_j[[t]][[j]]<-c(beta[j],beta_t_j[[t-1]][[j]]+1)
      cj<-rep(0,length(alpha_t_j[[t-1]][[j]]))
```

```
        for (i in 1:length(alpha_t_j[[t-1]][[j]])){
          integrand <- function(la) {dgamma(la,alpha_t_j[[t-1]][[j]][i],
                                  beta_t_j[[t-1]][[j]][i])*dpois(spikes[t-1],la)}
          cj[i]<-integrate(integrand, lower = 0, upper = Inf)[[1]]
        }
        c_t_j[[t]][[j]]<-cj
      }
    }
    w_t_j[[t]]<-as.list(rep(1/k,k))
    for (i in 1:(t-1)){
      wc_i<-sum(sapply(c_t_j[[t]], "[", i)*sapply(w_t_j[[t-1]], "[", i))
      for (j in 1:k){
        w_t_j[[t]][[j]]<-c(w_t_j[[t]][[j]],c_t_j[[t]][[j]][i]
                           *w_t_j[[t-1]][[j]][i]/wc_i)
      }
    }
  }
  return(w_t_j)
}
```

Listing 5: Determine the weights of a mixture distribution (Section 4.4) for each cycle k and each possible runlength. The BOCD in case of a mixture distribution can be obtained by adjusting the update process of Function 3.

```
BOCD_discrete_rate_plus_phase<-function(X,lam,phi,a0,b0,sigma,weights){
  # Input:
  #  X...       List of spike times: Each element of the list is a vector.
  #                                  Element k corresponds to spike times of cycle k.
  #  lam...     Vector of rate parameters.
  #  phi...     Vector of phase parameters.
  #  a0...      First shape parameter of the beta prior distr. for cp-probability.
  #  b0...      Second shape parameter of the beta prior distr. for cp-probability.
  #  sigma...   Standard deviation of the spike times.
  #  weights... Prior probabilities of each stimulus
  # Output:
  #  cp_pos...  Detected changepoint positions.
  T<-length(X)
  spikes<-unlist(lapply(X,length))
  runandX<-list() # P(r_t=i | X_(1:t))
  runandX[[1]]<-1
  rundandXandat<-list() # P(r_t=i , a_t=j | X_(1:t))
  rundandXandat[[1]]<-matrix(1,1,1)
  k<-length(phi) # number of stimuli
  w_t<-list() # update of probabilities of each stimulus
  w_t[[1]]<-matrix(weights,nrow = 1)
  for (t in 2:T){
    # Update of the posterior weights
    w_t[[t]]<-matrix(0,nrow=t,ncol=k)
    w_t[[t]][1,]<-rep(1/k,k)
    for (r in 2:t){
      if (spikes[t-1]>0){
        w<-w_t[[t-1]][r-1,]
        *dnorm(mean(X[[t-1]]),mean = phi,sd = sigma/sqrt(spikes[t-1]))
        *dpois(spikes[t-1],lambda = lam)
      }
      if (spikes[t-1]==0){
        w<-w_t[[t-1]][r-1,]*dpois(spikes[t-1],lambda = lam)
      }
      w_t[[t]][r,]<-w/sum(w)
    }
    # Calculate changepoint probability
    rtat<-matrix(0,nrow=t,ncol=t)
    if (spikes[t]>0){
      rtat[1,]<-c(0,colSums(rundandXandat[[t-1]])
                  *sum(dnorm(mean(X[[t]]),mean = phi,sd = sigma/sqrt(spikes[t]))
                    *dpois(spikes[t],lambda = lam))/k*(seq(2,t,1)-1+a0-1)/(t+a0+b0-2))
    }
    if (spikes[t]==0){
      rtat[1,]<-c(0,colSums(rundandXandat[[t-1]])*sum(dpois(0,lambda = lam))/k
                  *(seq(2,t,1)-1+a0-1)/(t+a0+b0-2))
    }
    # Calculate growth probabilities
    if (spikes[t]>0){
      rtat[2:t,-t]<-rundandXandat[[t-1]]*(matrix(t-seq(2,t,1)+b0,nrow=t-1,ncol=t-1,
                    byrow=T))/(t+a0+b0-2)*matrix(w_t[[t]][-1,]
                    %*%(dnorm(mean(X[[t]]),mean = phi,sd = sigma/sqrt(spikes[t]))
                    *dpois(spikes[t],lambda = lam)),nrow=t-1,ncol=t-1,byrow=F)
    }
    if (spikes[t]==0){
      rtat[2:t,-t]<-rundandXandat[[t-1]]
                    *(matrix(t-seq(2,t,1)+b0,nrow=t-1,ncol=t-1,byrow=T))/(t+a0+b0-2)
                    *matrix(w_t[[t]][-1,]%*%dpois(spikes[t],lambda = lam),
                                                nrow=t-1,ncol=t-1,byrow=F)
    }
    # Scaling for a robust calculation
    rundandXandat[[t]]<-rtat/sum(rtat)
  }
  rundandX<-lapply(rundandXandat,function(x){rowSums(x)})
  rundandX<-lapply(rundandX,unlist)
  # Determine changepoints (take the most likely runlength at the last cycle)
  cp_pos<-c()
  rt<-max(which(rundandX[[T]]==max(rundandX[[T]])))
  cp_pos<-c(T-rt+1,cp_pos)
  t<-T-rt
  while (t>0){
    rt<-max(which(rundandX[[t]]==max(rundandX[[t]])))
    cp_pos<-c(t-rt+1,cp_pos)
    t<-t-rt
  }
  cp_pos<-cp_pos[-1]
  return(cp_pos)
}
```

Listing 6: BOCD for a discrete stimuli setting (as used in Section 5), assuming a change in rate and phase simultaneously. Analog to Function 4 we can create an online version of the algorithm.

German Summary

In unserem Gehirn übertragen Neurone Information, indem sie elektrischen Entladungen, genannt Spikes, emittieren. Die Zeitpunkte, an denen Spikes auftreten, werden über die Zeit gemessen und als Spike Train bezeichnet (Figur 1 A, rote Striche). Von zentraler Bedeutung ist dabei die Identifikation und Bewertung verschiedener Signalkomponenten. Eine Möglichkeit Information zu kodieren besteht in der Variation der Anzahl emittierter Spikes, im Folgenden mit Rate bezeichnet, womit sich im Rahmen einer großen Neuronenpopulation Information genau übertragen lässt (Softky and Koch, 1993; Shadlen and Newsome, 1998; Pouget et al., 2000). Eine weitere Möglichkeit besteht in der zeitlich exakten Platzierung von Spikes, nachfolgend als genaue Phase bezeichnet, womit sich Information zusätzlich zur Rate übertragen lässt und die Robustheit gegenüber Fehlerrauschen erhöhen lässt (Nelken et al., 2005; Montemurro et al., 2008; Kayser et al., 2009; Cattani et al., 2015; Bieler et al., 2017).

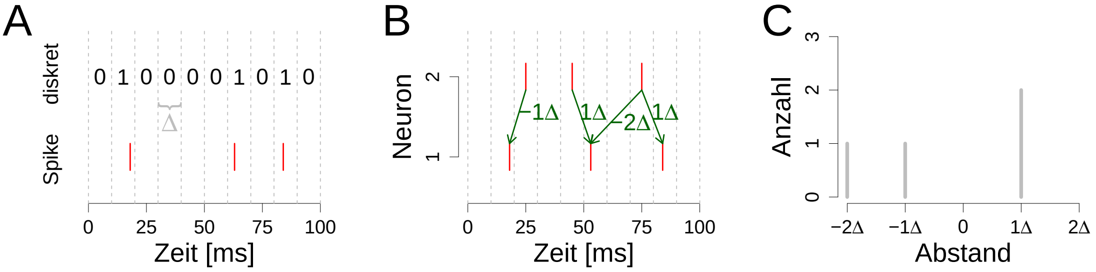


Abbildung 1: Schematische Darstellung eines CCH. A. Theoretische Spike Zeiten (rot) transformiert in eine diskrete Zeitreihe mit Auflösung Δ. B. Auftretende Spike Abstände bis zu einem maximalen Abstand von $D = 2\Delta$. C. Das resultierende CCH von Neuron 1 und Neuron 2 in B.

In vielen Fällen weisen neuronale Feuermuster jedoch große Variabilität auf. Dadurch lassen sich Phasen nur über einen langen Beobachtungszeitraum genau messen. In dieser Arbeit liegt der Fokus auf sogenannten ungenauen Phasen, die in synchronen Oszillationen von Neuronenpopulationen gemessen wurden. Üblicherweise wird dazu in der Praxis das Kreuzkorrelation Histogramm (CCH) betrachtet. Da sich die Feueraktivität nur mit einer gewissen Auflösung Δ messen lässt, liegt in der Praxis ein Spike Train in diskreter Form vor (Figur 1 A, 0 - 1). Für zwei Spike Trains verschiedener Neurone bestimmt man für jedes Vielfache $j \cdot \Delta$, $j \in \mathbb{Z}$, bis zu einem maximalen Abstand D, wie viele Spikes von Neuron 1 genau den Abstand $j \cdot \Delta$ zu einem Spike von Neuron 2 besitzen (Figur 1 B für $D = 2$). Die entsprechende Anzahl wird für jeden Abstand in einem Histogramm eingetragen (Figur 1 C). Ein repräsentatives CCH für simulierte Daten, geglättet mit einem Gaussian Kernel, ist in Figur 2 A für zwei simulierte Spike Trains mit einem Phasenunterschied von 2 ms gezeigt.

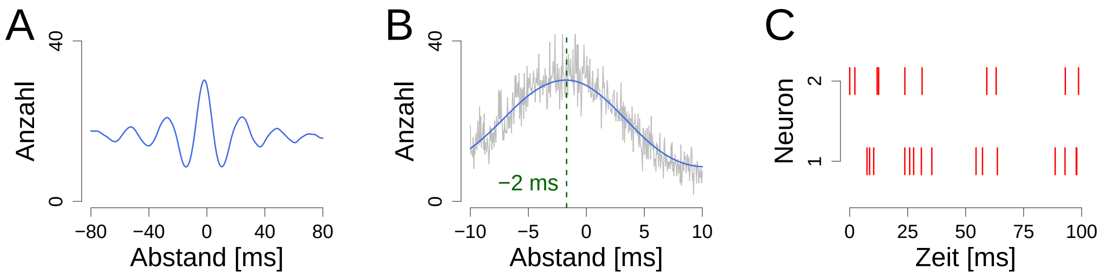


Abbildung 2: Wir betrachten zwei Spike Trains mit jeweils 1000 Oszillationszyklen (≈ 25 s) simuliert gemäß unseres GLO Models mit $\mu_B = 25$, $\sigma_B = 6$ und einer Spike Präzision von $\sigma = 4$. Neuron 1 hat eine höhere Rate von $\lambda^{(1)} = 4$ erwarteten Spikes je Zyklus, Neuron 2 emittiert in Erwartung $\lambda^{(2)} = 2$ Spikes. Die Phase von Neuron 1 ist $\varphi^{(1)} = 2$, die Phase von Neuron 2 ist $\varphi^{(2)} = 0$. A. CCH mit einem maximalen Abstand von $D = 80$ ms geglättet mit einem Gaussian Kernel, sd=1 ms. B. Hauptpeak des CCH : In grau die Rohwerte, in blau die Anzahlen geglättet mit einem Gaussian Kernel. Der vorherrschende Phasenabstand von 2 ms lässt sich mittels des vollständigen Spike Trains (25 s) genau messen, grün gestrichelte Linie. C. Ein Abschnitt von 100 ms der simulierten Spike Trains.

Welche Rolle ungenaue Phasen in der Informationsverarbeitung wahrnehmen, ist noch eine offene Frage. Einerseits wurden diese als Phänomen betrachtet, resultierend aus der ungenauen Arretierung der Feueraktivität an den Oszillationszyklus (Buzsáki and Chrobak, 1995; Roelfsema et al., 1997), und sie lassen sich auch nicht in kleinen Zeitfenstern, wie einem einzigen Oszillationszyklus identifizieren (Schneider and Nikolić, 2006) (vgl. Figur 2 C für vier Zyklen). Andererseits lassen sich ungenaue Phasen, die in akkumulierten Daten mit hoher Präzision gemessen werden können (vgl. Figur 2 B für einen Zoom in das CCH), nicht durch Zufall erklären (Schneider et al., 2006) und auch nicht, dass sich diese systematisch mit dem Stimulus ändern (Havenith et al., 2011). Daher bleibt unklar, ob und wie viel Information zusätzlich zur Rate durch ungenaue Phasen übertragen werden können. Die Fragestellung ist von besonderem Interesse, wenn man miteinbezieht, dass die Informationsverarbeitung im Gehirn sehr schnell abläuft und nur einzelne oder wenige Oszillationszyklen zur Verfügung stehen (Osram et al., 1999; Gautrais and Thorpe, 1998; Abeles, 1994).

Um dieser Fragestellung nachgehen zu können, betrachten wir eine modifizierte Version eines doppelt-stochastischen Spike Train Models, das die zeitlichen Feuermuster empirischer Spike Trains sowohl einzeln (Bingmer et al., 2011; Schiemann et al., 2012) als auch in Interaktion (Schneider and Nikolić, 2008) sehr gut erfassen konnte.

Das 'Gaussian Locking to a free Oscillator' (GLO) Model nimmt an, dass alle M Neurone denselben oszillatorischen Hintergrundrhythmus $\mathbb{B}$ teilen, der als stationäre Irrfahrt $(B_i)_{i\in\mathbb{Z}}$ mit einer Zuwachsverteilung $\mathcal{N}(\mu_B, \sigma_B)$ repräsentiert wird (vgl. Figur 3 für $M = 2$ Neurone). An jedem Beat wird für jedes Neuron $m \in \{1, \dots, M\}$ unabhängig eine Poisson-verteilte Anzahl Spikes $N_s^{(m)}$ mit Rate $\lambda_s^{(m)} \geq 0$ gewählt, wobei der Index s anzeigt, dass die Neurone auf Stimulus $s \in \{1, \dots, S\}$ reagieren. Anschließend werden unabhängige Normal-verteilte Spike Zeiten $X_{is}^{(m)}$, $i = 1, \dots, N_s^{(m)}$ mit Erwartungswert $\varphi_s^{(m)} \in \mathbb{R}$ und Varianz $\sigma^2 \geq 0$ gezogen und um den Beat platziert. Wir nehmen also an, dass die Präzision σ für alle Stimuli und Neurone gleich ist. Um einen Beat lässt sich damit die neu entstehende Feueraktivität eines

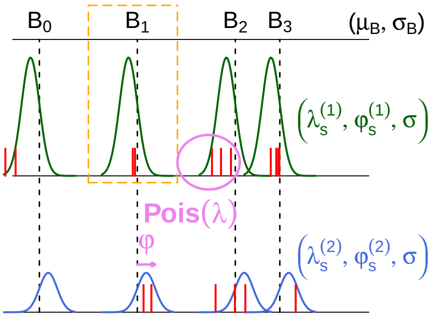


Abbildung 3: GLO Model für $M = 2$ Neurone, die auf Stimulus s reagieren. In grün bzw. blau ist die Feueraktivität von Neuron 1 bzw. Neuron 2 dargestellt.

Neuron m als inhomogener Poisson Prozess mit Intensität (Bingmer, 2012)

$$\rho_s^{(m)}(t) = \frac{\lambda_s^{(m)}}{\sqrt{2\pi}} \exp\left(-\frac{(\varphi_s^{(m)} - t)^2}{2}\right), s \in \{1, \ldots, S\},$$

beschreiben. Um die Bedeutung ungenauer Phasen zu erforschen, betrachten wir zwei konkrete Aufgaben in der Informationsverarbeitung: Erstens, können ungenaue Phasen helfen, den richtigen Stimulus $s \in \{1, \ldots, S\}$ zu erkennen, falls nur einer oder wenige Zyklen beobachtbar sind? Zweitens, können ungenaue Phasen helfen Änderungen im Stimulus zu detektieren, insbesondere wenn gefordert ist, schnell zu entscheiden? Dabei betrachten wir die Parameterbereiche, die in empirischen Spike Trains beobachtet wurden, welche $\lambda \in [0, 4]$ und $\varphi \in [0, 0.75]$ für $\sigma = 1$ sind (Havenith et al., 2011; Schneider, 2008; Schneider and Nikolić, 2006).

Stimulus korrekt erkennen (Kapitel 2)

Die Aufgabenstellung für einen Zyklus und $S = 2$ Stimuli ist in Figur 4 A für $M = 1$ Neuron und in B für $M = 2$ Neurone illustriert: Für jedes Neuron m kennen wir die Ratenparameter $\boldsymbol{\lambda}^{(m)} = (\lambda_1^{(m)}, \ldots, \lambda_S^{(m)})$ und Phasenparameter $\boldsymbol{\varphi}^{(m)} = (\varphi_1^{(m)}, \ldots, \varphi_S^{(m)})$ und müssen auf Basis der Spike Zeiten in einem Zyklus entscheiden, welcher Stimulus s den Neuronen präsentiert wurde. Wir nehmen dabei an, dass jeder Stimulus mit gleicher Wahrscheinlichkeit präsentiert wird und dass uns der Startpunkt des Zyklus bekannt ist. Als Entscheidungsregel verwenden wir die 'Bayesian Decision Rule' und wählen den Stimulus, der für die beobachtete Realisierung am Wahrscheinlichsten ist. Hierbei sind die Anzahl an Spikes n und die mittlere Spike Zeit $\bar{x}$ suffizient für Rate und Phase. Damit wird für ein Neuron der Beobachtungsraum $\mathbb{N} \times \mathbb{R}$ in S Akzeptanzregionen $A_1, \ldots, A_S$ eingeteilt, vgl. Figur 5 A für $S = 2$ Stimuli. Im Falle von $S = 2$ Stimuli lässt sich dieser angeben als

$$A_1 := \left\{ (n, \bar{x}) \middle| n \log \frac{\lambda_1}{\lambda_2} - \frac{\sqrt{n}}{\sigma}(\varphi_2 - \varphi_1) \left(\frac{\bar{x} - \varphi_1}{\sigma/\sqrt{n}} + \frac{\sqrt{n}}{\sigma} \frac{\varphi_1 - \varphi_2}{2} \right) > \lambda_1 - \lambda_2 \right\}.$$

Um die Bedeutung der Phase zu bewerten, betrachten wir die *Detektionswahrscheinlichkeit* p_D, die mittlere Wahrscheinlichkeit, dass für Stimulus s auch $(N_s, \bar{X}_s)$ in seinen Akzeptanzbereich

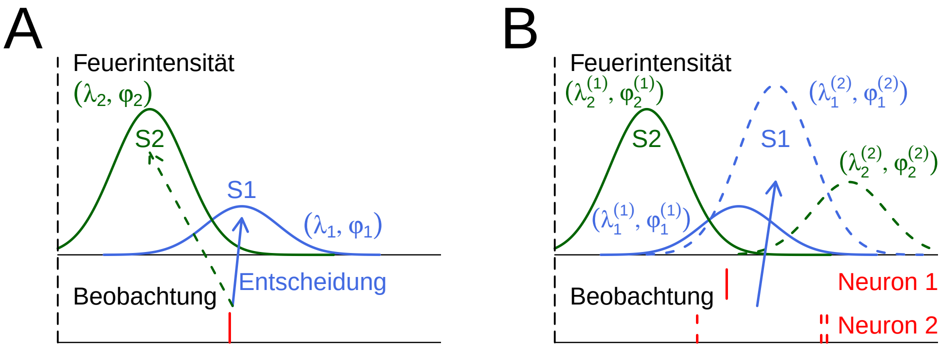


Abbildung 4: Entscheidungsaufgabe für einen Zyklus und $S = 2$ Stimuli und $M = 1$ Neuron (A) oder $M = 2$ Neurone (B).

A_s fällt. In Lemma 2.1.3 ist zu finden, wie sich p_D numerisch für ein Neuron berechnen lässt. Für zwei Neurone bestimmen wir p_D mittels Simulation (Lemma 2.2.2). Um die Bedeutung der Phase anhand von p_D quantifizieren zu können, bestimmen wir in Abhängigkeit der maximalen Rate λ_M und maximalen Phase φ_M zuerst optimale Ratenparameter, die p_D im Falle von identischen Phasen maximieren, und optimale Phasenparameter, für identische Raten, und vergleichen anschließend den Zuwachs in p_D für die optimale Kombination von Rate und Phase.

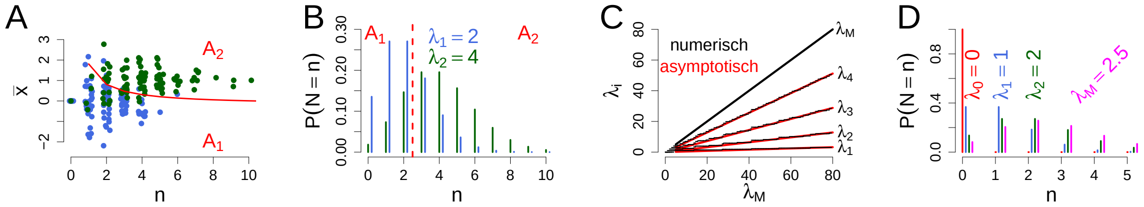


Abbildung 5: Akzeptanzregionen für zwei Stimuli im Falle von Raten- und Phasencodierung (A) und reiner Ratencodierung (B). C. Optimale Ratenparameter verglichen mit der asymptotischen Lösung. D. Optimale Ratenparameter für den Fall $S \geq \lambda_M$ ($S = 3$ Stimuli und $\lambda_M = 2.5$).

Im Falle reiner Ratencodierung vereinfachen sich die Akzeptanzregionen (siehe Figur 5 B), wodurch sich zum einen mittels dynamischer Programmierung die optimalen Raten numerisch bestimmen lassen (Figur 5 C, schwarze Stufenfunktion) und zum anderen die asymptotische ($\lambda_M \to \infty$) Relation (rote Linien)

$$\lambda_s = \left(\frac{s}{S}\right)^2 \lambda_M, \qquad s = 1, \dots, S,$$

zeigen lässt. Für kleine Raten $\lambda_M \leq S$ sind die optimalen Ratenparameter besonders leicht zu bestimmen, da an der Stelle $k \in \mathbb{Z}$ die $Pois(k)$-Verteilung maximal unter allen Poissonverteilungen ist, vgl. Figur 5 D. Im Falle von reiner Phasencodierung ist die äquidistante Platzierung der Phasenparameter von 0 bis φ_M auf Grund der Annahme gleicher Präzision σ und der Symmetrie der Normalverteilung optimal.

Komplizierter ist der Fall einer gleichzeitigen Codierung mit Rate und Phase. Hier bestimmen wir die optimalen Parameter numerisch. Die grundsätzliche Struktur der optimalen Parameter kann in Figur 6 A für $S = 2$ Stimuli beobachtet werden: Ist φ_M klein, codieren wir ausschließlich mit der Rate; für wachsendes φ_M wird mehr Information von der Rate auf die Phase übertragen, bis schließlich für $\varphi_M \approx 0.6$ eine reine Phasencodierung gewählt wird und keine Information mehr über die Rate übertragen wird. In Figur 6 B lässt sich erkennen, dass für $\varphi_M = 0.75$ durch die Hinzunahme der Phase (grün) nennenswert mehr Information übertragen lässt, als mit einer reinen Ratencodierung (rot). Dies gilt aber nicht mehr für $M = 2$ Neurone und

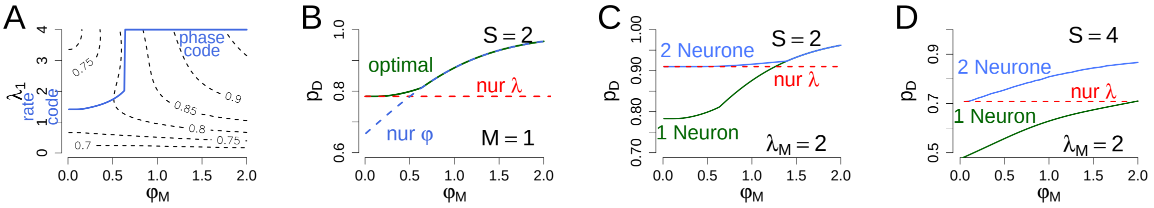


Abbildung 6: A. Optimale Rate λ_1, die p_D maximiert, für $\lambda_M = 4$ als Funktion von φ_M. B. Maximale p_D für $\lambda_M = 4$, $S = 2$ Stimuli und $M = 1$ Neuron, auf Basis von λ und φ (grün) in Vergleich zu einem reinen Ratencode (rot) und einem reinen Phasencode (blau). C (D) Maximale p_D für $\lambda_M = 2$, $S = 2$ ($S = 4$) Stimuli und $M = 2$ Neurone, auf Basis von λ und φ und zwei Neurone (blau), ein Neuron $\lambda_M = 4$ (grün) im Vergleich zu einem reinen Ratencode (rot) für zwei Neurone.

$S = 2$ Stimuli, siehe Figur 6 C blaue versus rote Linie. Der Grund liegt darin, dass allein mit der Rate Information Binär sehr stabil codiert werden kann: Bei einem Stimulus feuert nur Neuron 1, bei dem anderen nur Neuron 2 und bei einem dritten feuern beide. Deshalb werden mindestens $S \geq 2^M$ Stimuli benötigt, um einen deutlichen Anstieg in p_D im Vergleich zur Ratencodierung beobachten zu können, siehe Figur 6 D für $S = 4$ Stimuli. Festzuhalten ist, dass sich mit zwei Neuronen (jeweils maximal λ_M mögliche Spikes) deutlich mehr Information übertragen lässt als mit einem Neuron (maximal $2\lambda_M$ mögliche Spikes).

Grundsätzlich bleiben die Ergebnisse auch für zwei Oszillationszyklen bestehen. Jedoch benötigen wir hier genauere Phasen, um denselben Anstieg in p_D, verglichen zu einer reinen Ratencodierung, zu beobachten. Dies ist der zusätzlichen Unsicherheit geschuldet, dass nicht vorgeben ist, welchem Oszillationszyklus jeder einzelne Spike zuzuordnen ist.

Weiterhin sind die Akzeptanzbereiche stabil gegenüber der gewählten Klassifizierung-Technik Dazu haben wir unsere Ergebnisse basierend auf der 'Bayesian Decision Rule' mit der 'Linearen Diskriminanzanalyse' verglichen.

Auch untersuchen wir ein alternatives Maß zu p_D, nämlich den globalen Detektionsfehler e_D, der die Beziehung der Stimuli untereinander miteinbezieht. Dies erhöht deutlich den Rechenaufwand, resultiert aber in leichter identifizierbaren optimalen Raten und Phasenparametern, da keine Stimuli mit gleichzeitig mittlerem Raten- und mittlerem Phasenparameter als optimal auftreten.

Stimulusänderungen korrekt detektieren (Kapitel 4)

Um die Fragestellung zu untersuchen, ob ungenaue Phasen helfen können Änderungen im Stimulus zu detektieren, betrachten wir folgendes Change Point Modell (illustriert in Figur

7), dass die Feueraktivität aus dem GLO Modell verwendet. Erstens nehmen wir an, dass Change Points unabhängig und mit gleicher Wahrscheinlichkeit η in jedem Zyklus auftreten. Dazu sei $Y_1, Y_2, \ldots$ eine Folge unabhängiger Bernoulli Zufallsvariablen, wobei Y_k angibt, ob zwischen Zyklus $k-1$ und k ein Change Point ist. Zweitens, sei $\Lambda_0, \Lambda_1, \ldots$ eine Folge von u.i.v. Ratenparametern mit Prior Verteilung $\pi_\lambda(\cdot)$ und $\lambda_0, \lambda_1, \ldots$ eine zufällige Realisierung. Analog sei $\Phi_0, \Phi_1, \ldots$ eine Folge u.i.v. Phasenparametern mit Prior Verteilung $\pi_\varphi(\cdot)$ und $\varphi_0, \varphi_1, \ldots$ eine zufällige Realisierung. An jedem Change Point wird eine neue Realisierung von Λ, Φ gezogen, wir nehmen also an, dass Change Points stets simultan in Rate und Phase auftreten. Bezeichne dazu $A_k := \sum_{i=1}^{k} Y_i$ die Anzahl Change Points bis zur Zeit k, wobei wir $A_0 := 0$ setzen. Dann werden in Zyklus k insgesamt $N_k \sim Pois(\lambda_{A_k})$ Spikes gewählt und unabhängig gemäß einer $\mathcal{N}(\varphi_{A_k}, \sigma^2)$-Verteilung platziert, wobei wir annehmen, dass die Präzision σ der Spike Zeiten fest und bekannt ist. Wieder sei $\sigma = 1$ und wir skalieren die Phasenparameter entsprechend.

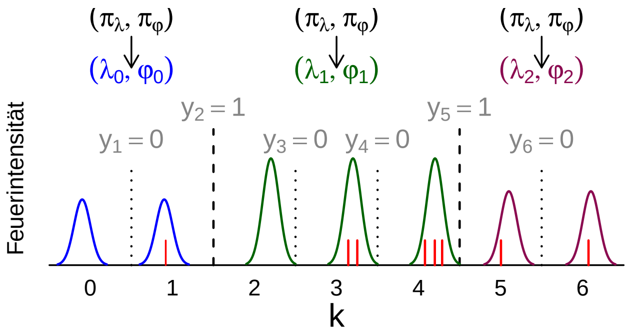


Abbildung 7: Change Point Modell. In einem Zyklus folgt die Feuerintensität einem inhomogenen Poisson Prozess mit Rate λ und Phase φ wie im GLO Modell.

Zur Detektion von Change Points verwenden wir einen Bayesian Online Change Point Algorithmus (BOCD) (Adams and MacKay, 2007), welcher die Schätzung der Change Point Wahrscheinlichkeit η beinhaltet (Wilson et al., 2010). Der Algorithmus kann die Modellannahmen exakt erfassen und bestimmt, gegeben die Annahmen, die Wahrscheinlichkeit für einen Change Point zu jedem Zeitpunkt k exakt. Er bezieht Informationen über neurophysiologisch relevante Parameterbereiche und Verteilungsannahmen mit ein. Dadurch kann der BOCD auf kurze Zeitreihen angewandt werden und ermöglicht eine Untersuchung von Rate und Phase auf Basis von wenigen Zyklen, im Gegensatz zu asymptotischen Verfahren.
Kernidee des BOCD ist die Betrachtung der Verteilung der Runlänge, welche die Zeit bis zum letzten Change Point angibt, formal definiert als

$$R_k := \begin{cases} \min(i \geq 0: A_k - A_{k-1-i} = 1), & \text{falls } A_k > 0, \\ k, & \text{sonst.} \end{cases}$$

Mittels des Satzes von Bayes und der Annahmen des Change Point Modells, lässt sich $\mathbb{P}(R_k = r_k \,|\, X_{0:k} = x_{0:k}) \; \forall r_k \in 0, \ldots, k$, wobei $x_{0:k}$ die Beobachtungen in Zyklen 0 bis k bezeichnen, rekursiv bestimmen, vgl. Figur 8 A mit verschiedenen Grautönen codierte Punkte. Um die Predictive Wahrscheinlichkeit eines Change Points berechnen zu können, nehmen wir an, dass auch diese uniform aus $[0, 1]$ gezogen wird, vgl. Figur 8 B. Zur Detektion von Change Points betrachten wir bei einer gegebenen Zeitreihe der Länge K die Verteilung der Runlänge

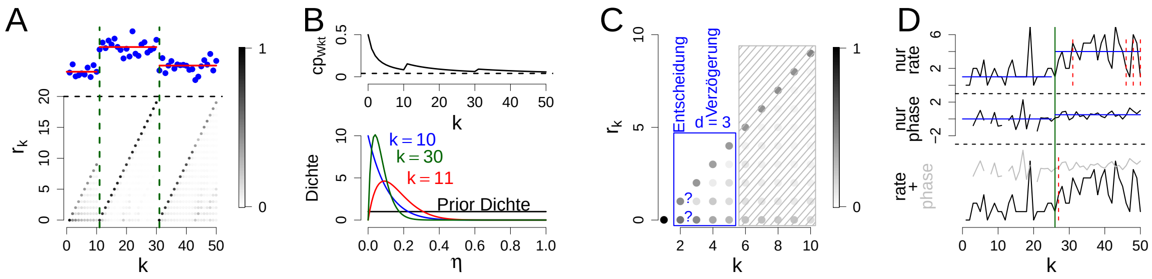


Abbildung 8: A. Eine Beispielfolge von normalverteilten Zufallsvariablen mit zwei Change Points im Erwartungswert und die geschätzte Runlängenverteilung für jeden Zyklus k, illustriert mit einem Graucode. In grün die geschätzten Change Points mit maximaler Posterior Wahrscheinlichkeit zur Zeit $K = 50$. B. Illustration der Predictive Change Point Wahrscheinlichkeit $cp_{prob} = \mathbb{P}(Y_k = 1 \mid A_{k-1} = a_{k-1})$ für den Change Point Pfad geschätzt in A. C. BOCD mit online Entscheidung, Verzögerung $d = 3$. D. Beispielfolge von Zyklen mit einem Change Point in der Mitte (grün) und den mit dem BOCD detektierten Change Points in rot separat für die drei Analysearten.

R_K und wählen die Runlänge r_K mit maximaler Wahrscheinlichkeit und fahren analog fort rückwärts in der Zeit. Da die Detektion von Change Points nicht als online angesehen werden kann, sondern nur das Updaten der Verteilung der Runlänge, betrachten wir außerdem einen modifizierten Algorithmus (BOCD mit online Entscheidung), der annähernd sofort mit einer kurzen Verzögerung entscheidet (Figur 8 C).

Um die Bedeutung von Rate und Phase in der Detektion von Change Points zu bewerten, wenden wir erstens den BOCD nur auf die Anzahl Spikes an (nur Rate, rot), oder zweitens nur auf die Spikezeiten (nur Phase, blau) oder drittens gleichzeitig auf Anzahl und Zeitpunkte (Rate und Phase, grün). Dabei konzentrieren wir uns auf Zeitreihen mit einem Change Point genau in der Mitte, an dem wir Sprünge in Rate und Phase betrachten, die plausibel für den gegebenen Parameterbereich sind, für ein Beispiel siehe (Figur 8 D).

Grundsätzlich hat sich ergeben, dass nur die Phase deutlich weniger Change Points korrekt (der wahre Change Point ist höchstens drei Zyklen entfernt) detektieren kann als nur die Rate mit und ohne online Entscheidung (Figur 9 A und C). Nur die Rate hingegen detektiert sehr viele Change Points, die gar nicht vorhanden sind (Figur 9 B und D). Betrachtet man Rate und Phase gemeinsam, lassen sich mehr Change Points korrekt detektieren und die Anzahl falsch detektierter Change Points wird deutlich reduziert. Das ist insbesondere interessant für den BOCD mit online Entscheidung, da hier die Hinzunahme von ungenauen Phasen es erst ermöglichen, schnell und fehlerfrei (mit knapp über fünf Zyklen Verzögerung) Änderungen im Stimulus zu detektieren.

In der vorangegangenen Betrachtung verwenden wir konjugierte Prior Verteilungen, die im relevanten Parameterbereich alle Raten und Phasen möglichst gleich gewichten. Möglicherweise haben wir bzw. unser Gehirn in manchen Situationen spezielle Information, wie z.B. zwei Typen von Ratenparametern oder kleine Raten treten nur mit kleinen Phasen auf. Durch Konvexkombination von konjugierten Verteilungen lässt sich derartige Information effizient berücksichtigen. Es zeigt sich, dass auch in diesem Fall mittels der Phase zusätzlich Information übertragen werden kann und die Phase sogar auf gleicher Augenhöhe mit der Rate agieren

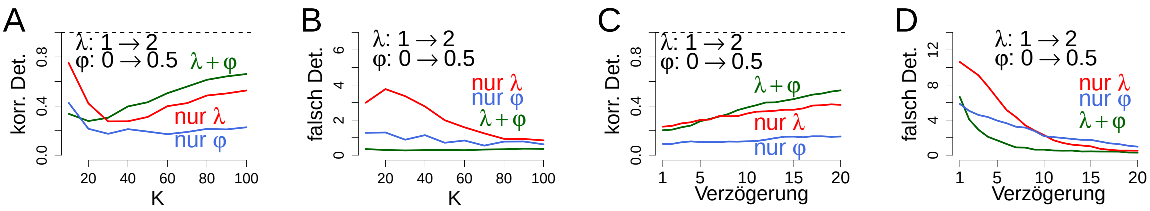


Abbildung 9: Anwendung des BOCD (A und B) und des BOCD mit online Entscheidung (C und D) für Zeitreihen mit einem Change Point in der Mitte. Die Phase springt von 0 auf 0.5, die Rate von 1 auf 2. Mittlere Anzahl korrekt detektierter Change Points (A,C) und mittlere Anzahl falsch detektierter Change Points (B,D) für Folgen von Zyklen der Länge $K = 10$ bis $K = 100$ (BOCD) bzw. der Länge $K = 100$ (BOCD mit online Entscheidung).

kann, da die Genauigkeit der Prior Information eine bessere Unterscheidung ermöglicht, auch von kleinen Phasenunterschieden.

Weiterhin sind wir davon ausgegangen, dass die Präzision σ bekannt und fest ist. Ändert sich diese, aber wir nehmen diese fälschlicherweise als konstant hat, beobachten wir eine systematische Überschätzung an Change Points in der Phase. Dies lässt sich verhindern, indem wir auch die Präzision als unbekannt und zufällig modellieren. Auch in diesem Fall können ungenaue Phasen die Anzahl korrekt detektierter Change Point erhöhen und gleichzeitig die falschen Detektionen reduzieren.

Empirische Neurone (Kapitel 5)

Zum Abschluss geben wir noch einen Ausblick, wie sich unser Modellierung eignet, verborgene Kodierungsmechanismen in Daten ausfindig zu machen. Dazu betrachten wir empirische Neurone aus Havenith et al. (2011). Die Autoren haben das Antwortverhalten von acht Neuronen auf 12 Stimuli - sich bewegende Balken, wobei sich deren Bewegungsrichtung um 30° Schritte unterscheidet, vgl. Figur 10 A - gemessen.

Wir konnten beobachten, dass man im Sinne der Detektionswahrscheinlichkeit auch in empirischen Neuronen annähernd optimale Parameterkombinationen beobachten kann (Neurone einzeln betrachtet für einen Oszillationszyklus). Etwa die Hälfte der Neurone zeigt jedoch sehr kleine Raten innerhalb eines Zyklus, weshalb ein Zugewinn in p_D durch Hinzunahme der Phase nicht möglich ist. Es lässt sich aber beobachten, dass die Fähigkeit Change Points (Änderungen in der Orientierung der Balken) korrekt detektieren zu können, durch Hinzunahme der Phase wächst, insbesondere wenn man mehrere Neurone mit denselben Raten- und Phasenparametern annimmt.

Betrachtet man alle Neurone gleichzeitig kann kaum eine Erhöhung der Detektionswahrscheinlichkeit (0.54 zu 0.56) beobachtet werden, was zum einen an der geringen Anzahl Stimuli im Verhältnis zu der Anzahl Neurone liegt und zum anderen durch die niedrigen Ratenparameter verstärkt wird. Betrachtet man insbesondere genau die auftretenden Fehlentscheidungen, stellt man fest, dass dies größtenteils direkt benachbarte Stimuli betrifft (Figur 10 B). Bereits nach einem Zyklus lässt sich mit den gemessenen Neuronen mit hoher Wahrscheinlichkeit die Orientierung des Balken mit einer Genauigkeit von ±30° bestimmen. Dieses Phänomen überträgt sich auf die Detektion von Change Points: Auch hier ist es besonders schwierig kleine Änderungen in der Orientierung (±30°) zu detektieren, wobei Änderungen von mindesten

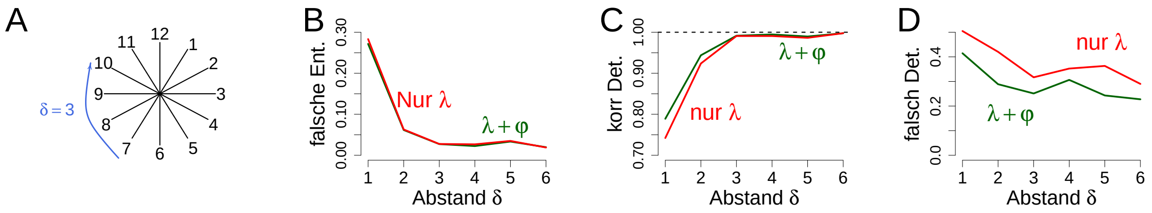


Abbildung 10: Detektionswahrscheinlichkeit und Change Point Detektion in Abhängigkeit des Stimulusabstandes. A. Illustration des Abstandes für die zwölf gemessenen Stimuli. B. Mittlere Wahrscheinlichkeit Stimuli mit dem Abstand δ zu verwechseln. C-D. Ergebnisse des BOCD mit online Entscheidung und einer fixen Verzögerung von $d = 5$ für Spike Trains der Länge $K = 100$ mit einem Change Point in der Mitte bei $K/2 = 50$. C. (D.). Mittlere Anzahl von korrekt (falsch) detektierten Change Points in Abhängigkeit des Abstandes des Stimulus vor und nach dem Change Point.

90° fast immer erkannt werden (Figur 10 C). Die Phase ermöglicht dabei gerade bei sehr ähnlichen Stimuli eine verbesserte Detektion und reduziert insgesamt die Anzahl an falschen Detektionen (Figur 10 D).
Zusammenfassend können wir festhalten, dass die einfache und direkte Beschreibung der Rate und Phase in dem stochastischen Modell ergeben hat, dass die Verwendung von ungenauen Phasen nicht nur die Wahrscheinlichkeit erhöhen kann, den korrekten Stimulus zu detektieren, sondern auch die Anzahl korrekt detektierter Änderungen im Stimulus erhöht. Außerdem wird die Robustheit erhöht, bzw. die Häufigkeit von Fehlmeldungen hinsichtlich von Änderungen reduziert. Ferner lassen sich mit unserem Modell Kodierungsprinzipien in empirischen Aufzeichnungen erforschen. Beispielsweise in empirischen Parametern, gewonnen aus Havenith et al. (2011), konnten wir feststellen, dass für deutlich unterschiedliche Stimuli bereits ein Oszillationszyklus ausreichend ist, um diese sehr sicher zu unterscheiden, sogar allein mit der Rate. Änderungen zwischen sehr ähnlichen Stimuli lassen sich hingegen besser durch Hinzunahme der Phase erkennen. Diese Ergebnisse suggerieren, dass kleine bzw. ungenaue Phasen zur Informationsverarbeitung beitragen können, indem sie die Wahrscheinlichkeit und Präzision in der korrekten Stimulusdetektion erhöhen und gleichzeitig eine robuste Detektion von Änderungen ermöglichen.

List of Figures

List of Notations and Abbreviations

Abbreviations

BOCD	Bayesian online change point detection algorithm
CCF	cross correlation function
CCH	cross correlation histogram
GLO	Gaussian locking to a free oscillator
LDA	linear discriminant analysis

Notations

General

M	number of neurons
S	number of stimuli
$\mathbb{B}$	background oscillation in the GLO process
μ_B	mean increment of the background oscillation $\mathbb{B}$
σ_B^2	variance of the increments of the background oscillation $\mathbb{B}$
$\Gamma(\cdot)$	Gamma function

Distributions

$Ber(\eta)$	Bernoulli distribution with success probability η
$Beta(a, b)$	Beta distribution with shape parameters a and b
$Binom(k, \eta)$	Binomial distribution with k trials and success probability η
$Exp(\lambda)$	Exponential distribution with rate λ

$Gamma(\alpha, \beta)$	Gamma distribution with shape parameter α and rate parameter β
$\mathcal{IG}(\gamma, \delta)$	Inverse-Gamma distribution with shape parameter γ and scale parameter δ
$\mathcal{NB}(k, \eta)$	Negative-binomial distribution with k successes and success probability η
$\mathcal{N}\left(\varphi, \sigma^2\right)$	Normal distribution with mean φ and variance σ^2
ϕ_{φ,σ^2}	density of normal distribution with mean φ and variance σ^2
$Pois(\lambda)$	Poisson distribution with rate λ
$\mathcal{T}_\nu(\mu, \chi^2)$	t-distribution with location parameter μ, scale parameter χ and ν degrees of freedom
$Unif(a, b)$	uniform distribution on the interval $[a, b]$
$Unif(\{a_1, \dots, a_n\})$	(discrete) uniform distribution on the elements $a_1, \dots, a_n$

Stimulus Encoding

$\boldsymbol{\lambda}^{(m)} = \left(\lambda_1^{(m)}, \dots, \lambda_S^{(m)}\right)$	rate parameters of neuron m
$\lambda_M^{(m)}$	maximal rate parameter of neuron m
$N_s^{(m)}$	$Pois\left(\lambda_s^{(m)}\right)$-distributed number of spikes
$\boldsymbol{\varphi}^{(m)} = \left(\varphi_1^{(m)}, \dots, \varphi_S^{(m)}\right)$	phase parameters of neuron m
$\varphi_M^{(m)}$	maximal phase parameter of neuron m
$X_{1s}^{(m)}, \dots, X_{n_s s}^{(m)}$	$\mathcal{N}\left(\varphi_s^{(m)}, \sigma^2\right)$-distributed independent spike times of n_s spikes
$\bar{X}_s^{(m)}$	$\mathcal{N}\left(\varphi_s^{(m)}, \sigma^2/n_s^{(m)}\right)$-distributed mean spike time
σ	precision of the spike times (mostly set to 1)
$\rho_s^{(m)}(\cdot)$	firing intensity of neuron m responding to stimulus s
$A_1, \dots, A_S$	acceptance regions of stimuli $1, \dots, S$
p_s	probability to detect stimulus s correctly if it is present
p_D	(average) detection probability
$p_{s_1 s_2}$	probability to falsely detect stimulus s_2 instead of correct stimulus s_1
$p^{(\delta)}$	average probability to misclassify two stimuli with distance of δ
e_D	(average) detection error weighted with the distance of stimuli

Bayesian Inference

$\mathcal{B}(\Pi, \mathcal{P})$	Bayesian model with prior distribution Π and $\mathcal{P}$ a family of sampling distributions
Θ	random variable with distribution Π
$X_1, \ldots, X_k$	random variables with conditional distribution P_θ given $\{\Theta = \theta\}$
$\pi(\cdot)$	prior density of Θ
$\pi(\cdot \mid X_{1:k} = x_{1:k})$	posterior density of Θ after observing $x_1, \ldots, x_k$
$p_\theta(\cdot)$	conditional weights or density of X
$p(\cdot)$	prior predictive distribution of X
$p(\cdot \mid X_{1:k} = x_{1:k})$	predictive distribution of X after observing $x_1, \ldots, x_k$
$\sim$	'distributed as' or 'proportional to'
k_0, t_0	prior parameters of the standard conjugate prior distribution of an exponential family distribution
$t(\cdot)$	sufficient statistic of θ for an exponential family distribution

Change point model

K	total number of cycles
k	current cycle
$Y_1, Y_2, \ldots$	background change point process ($Y_k \mid \{H = \eta\} \sim Ber(\eta)$)
H	random change point probability ($H \sim Beta(a_0, b_0)$)
A_k	number of change points up to time k ($A_k = \sum_{i=1}^k Y_i$)
R_k	random run length in cycle k (time since last change point)
r_k	realization of random run length in cycle k
$\Lambda_0, \Lambda_1, \ldots$	i.i.d. random rate parameters ($\Lambda \sim Gamma(\alpha_0, \beta_0)$)
$\lambda_0, \lambda_1, \ldots$	realizations of random rate parameters
N_k	Poisson distributed number of spikes in cycle k
$\Phi_0, \Phi_1, \ldots$	i.i.d. random phase parameters ($\Phi \sim \mathcal{N}(\mu_0, \tau_0^2)$ or $\Phi \sim \mathcal{N}(\mu_0, \varsigma^2/k_0)$ for random spike precision)
$\varphi_0, \varphi_1, \ldots$	realizations of random phase parameters
$X_{1:n_k}^{(k)}$	normally distributed spike times in cycle k
$\bar{X}_k$	mean spike time in cycle k

$\varsigma_0, \varsigma_1, \dots$	i.i.d. random precision parameters ($\varsigma \sim \mathcal{IG}(\gamma_0, (\gamma_0 - 1)\sigma_0^2)$, $\gamma_0 := (k_0 + 3)/2$)
$\tilde{\sigma}_0, \tilde{\sigma}_1, \dots$	realizations of random precision parameters

Bibliography

Abeles, M. (1994). *Models of Neural Networks II.* Springer, New York.

Adams, R. and MacKay, D. (2007). Bayesian online changepoint detection. *Technical Report. University of Cambridge, UK.* `https://arxiv.org/abs/0710.3742`.

Adrian, E. D. (1928). *The Basis of Sensations.* Norton, New York.

Alippi, C., Boracchi, G., Carrera, D., and Roveri, M. (2016). Change detection in multivariate datastreams: Likelihood and detectability loss. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2:1368–1374.

Bieler, M., Sieben, K., Cichon, N., Schildt, S., and Hanganu-Opatz, B. R. I. (2017). Rate and temporal coding convey multisensory information in primary sensory cortices. *eNeuro*, 4(2):ENEURO.003717.2017.

Bingmer, M. (2012). *A stochastic model for the joint evaluation of burstiness and regularity in oscillatory spike trains.* PhD thesis, Johann Wolfgang Goethe University.

Bingmer, M., Schiemann, J., Roeper, J., and Schneider, G. (2011). Measuring burstiness and regularity in oscillatory spike-trains. *Journal of Neuroscience Methods*, 201(2):426–437.

Bizley, J. K., Walker, K. M., King, A. J., and Schnupp, J. W. (2010). Neural ensemble codes for stimulus periodicity in auditory cortex. *Journal of Neuroscience Methods*, 30(14):5078–5091.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, volume 9. Institute of Mathematical Statistics.

Buzsáki, G. and Chrobak, J. (1995). Temporal structure in spatially organized neuronal ensembles: a role for interneuronal networks. *Current Opinion in Neurobiology*, 5(4):504–510.

Buzsáki, G. and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, 304(5679):1926–1929.

Camastra, F. and Vinciarelli, A. (2015). *Machine Learning for Audio, Image and Video Analysis - Theory and Applications.* Springer, London, second edition.

Cattani, A., Einevoll, G., and Panzeri, S. (2015). Phase-of-firing code. *Encyclopedia of Computational Neuroscience.* `https://arxiv.org/abs/1504.03954`.

Consonni, G. and Veronese, P. (1992). Conjugate priors for exponential families having quadratic variance functions. *Journal of the American Statistical Association*, 87(420):1123–1127.

Cox, D. R. (1962). *Renewal Theory.* Methuens monographs on applied probability and statistics. Methuen & CO LTD, London.

Cox, D. R. and Isham, V. (1980). *Point Processes.* CRC Monographs on Statistics & Applied Probability.

Daley, D. J. and Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes.* Springer Berlin.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281.

Diaconis, P. and Ylvisaker, D. (1985). *Quantifying prior opinion.* Bayesian statistics 2, Amsterdam: North-Holland, 133-156.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics.*

Fries, P. (2009). Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual Review of Neuroscience*, 32:209–224.

Gautrais, J. and Thorpe, S. (1998). Rate coding versus temporal order coding: a theoretical approach. *Biosystems Volume*, 48(1-3):57–65.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2013). *Bayesian Data Analysis.* Chapman & Hall/CRC, Boca Raton, third edition.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics.* Springer-Verlag New York.

Gutiérrez-Pena, E. and Smith, A. (1995). Conjugate parametrizations for natural exponential families. *Journal of the American Statistical Association, 90, 13471356.*

Gutiérrez-Pena, E. and Smith, A. (2003). Reference priors for exponential families. *Journal of Statistical Planning and Inference*, 110(1-2):35–54.

Havenith, M. N., Yu, S., Biederlack, J., Chen, N., Singer, W., and Nikolić, D. (2011). Synchrony makes neurons fire in sequence, and stimulus properties determine who is ahead. *The Journal of Neuroscience*, 31(23):8570–8584.

Kayser, C., Montemurro, M. A., Logothetis, N. K., and Panzeri, S. (2009). Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron*, 61(4):597–608.

Kersting, G. and Wakolbinger, A. (2010). *Elementare Stochastik.* Birkhäuser Basel.

Khazipov, R. and Luhmann, H. J. (2006). Early patterns of electrical activity in the developing cerebral cortex of humans and rodents. *Trends in Neurosciences*, 29(7):414–418.

Kingman, J. F. C. (1993). *Poisson Processes.* Oxford University Press, USA.

König., P., Engel, A., Roelfsema, P., and Singer, W. (1995). How precise is neuronal synchronization? *Neural Computation*, 7(3):469–485.

Liemant, A., Matthes, K., and Wakolbinger, A. (1988). *Equilibrium distributions of branching processes*, volume 34 of Mathematics and its Applications (East European Series). Kluwer Academic Publishers Group, Dordrecht,.

Lorenzo, P. M. D., Chen, J. Y., and Victor, J. (2009). Quality time: representation of a multidimensional sensory domain through temporal coding. *Journal of Neuroscience*, 29(29):9227–9238.

Mardia, K. V., Kent, J., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.

Montemurro, M. A., Panzeri, S., Maravall, M., Alenda, A., Bale, M. R., Brambilla, M., and Petersen, R. S. (2007). Role of precise spike timing in coding of dynamic vibrissa stimuli in somatosensory thalamus. *Journal of Neurophysiologyl*, 98(4):1871–1882.

Montemurro, M. A., Rasch, M. J., Murayama, Y., Logothetis, N. K., and Panzeri, S. (2008). Phase-of-firing coding of natural visual stimuli in primary visual cortex. *Current Biology*, 18(5):375–380.

Moore, G. P., Perke., D. H., and Segundo, J. P. (1966). Statistical analysis and functional interpretation of neuronal spike data. *Annual Review of Physiology*, 28:493–522.

Nelken, I., Chechik, G., Mrsic-Flogel, T., King, A., and Schnupp, J. (2005). Encoding stimulus information by spike numbers and mean response time in primary auditory cortex. *Journal of Computational Neuroscience*, 19(2):199–221.

Nemenman, I., Lewen, G. D., Bialek, W., and de Ruyter van Steveninck, R. R. (2008). Neural coding of natural stimuli: information at sub-millisecond resolution. *PLOS Computational Biology*, 4(3):e1000025.

Osram, M. W., Wiener, M. C., Lestienne, R., and Richmond, B. J. (1999). Stochastic nature of precisely timed spike patterns in visual system neuronal responses. *Journal of Neurophysiology*, 81(6):3021–3033.

Perkel, D. H., Gerstein, G. L., and Moore, G. P. (1967). Neuronal spike trains and stochastic point processes ii. simultaneous spike trains. *Biophysical Journal*, 7(4):419–440.

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132.

Roelfsema, P., Engel, A., Knig, P., and Singer, W. (1997). Visuomotor integration is associated with zero time-lag synchronization among cortical areas. *Nature*, 385(6612):157–161.

Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag New York.

Schiemann, J., Klose, V., Schlaudraff, F., Bingmer, M., Seino, S., Magill, P. J., Schneider, G., Liss, B., and Roeper, J. (2012). K-atp channels control in vivo burst firing of dopamine neuron in the medical substantia nigra and novelty-induced behavior. *Nature Neuroscience*, 15(9):1272–1280.

Schneider, G. (2008). Messages of oscillatory correlograms - a spike-train model. *Neural Computation*, 20(5):1211–1238.

Schneider, G., Havenith, M. N., and Nikolić, D. (2006). Spatio-temporal structure in large neuronal networks detected from cross correlation. *Neural Computation*, 18(10):2387–2413.

Schneider, G. and Nikolić, D. (2006). Detection and assessment of near-zero delays in neuronal spiking activity. *Journal of Neuroscience Methods*, 152(1-2):97–106.

Schneider, G. and Nikolić, D. (2008). A stochastic framework for the quantification of synchronous oscillation in neuronal networks. *Proceedings of the Fifth International Workshop on Computational Systems Biology, WCSB*, pages 169–172.

Shadlen, M. and Newsome, W. (1998). The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of Neuroscience*, 18(10):3870–3896.

Sherrington, C. S. (1933). The brain and its mechanisms. *American Journal of Sociology*, 40(5):705–705.

Singer, W. (1995). Development and plasticity of cortical processing architectures. *Science*, 270(5237):758–764.

Softky, W. and Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps. *Journal of Neuroscience*, 13(1):334–350.

Spitzer, F. (1976). *Principles of Random Walk.* Springer-Verlag, New York, Heidelberg, Berlin.

Srivastava, K. H., Holmes, C. M., Vellema, M., Pack, A. R., Elemans, C. P. H., Nemenman, I., and Sober, S. J. (2017). Motor control by precisely timed spike patterns. *Proceedings of the National Academy of Sciences*, 114(5):1171–1176.

Thompson, W. A. (1988). *Point Process Models with Applications to Safety and Reliability.* Chapman & Hall, London/New York,.

Thorpe, S. J., Delorme, A., and VanRullen, R. (2001). Spike based strategies for rapid processing. *Neural Networks*, 14(6-7):715–726.

Uhlhaas., P. J., Roux, F., Rodriguez, E., Rotarska-Jagiela, A., and Singer, W. (2009). Neural synchrony and the development of cortical networks. *Trends in Cognitive Sciences*, 14(2):72–80.

Vinck, M., Lima, B., Womelsdorf, T., Oostenveld, R., Singer, W., Neuenschwander, S., and Fries, P. (2010). Gamma-phase shifting in awake monkey visual cortex. *Journal of Neuroscience*, 30(4):1250–1257.

Wilson, R. C., Nassar, M. R., and Gold, J. (2010). Bayesian on-line learning of the hazard rate in change-point problems. *Neural Computation*, 22(9):2452–2476.

Woodroofe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis.* Society for Industrial and Applied Mathematics, Philadelphia.

Zimek, A., Schubert, E., and Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387.