Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

Cluster Regularization via a Hierarchical Feature Regression

Johann Pfitzinger[a,b,1]

[a] Goethe University Frankfurt Germany
[b] Stellenbosch University South Africa

ARTICLE INFO

Article history:
Received 12 February 2022
Revised 2 January 2024
Accepted 9 January 2024
Available online xxx

JEL classification:
C13
C53
C55
O47

Keywords:
Regularized regression
Group shrinkage
Machine learning
Supervised hierarchical clustering

ABSTRACT

The hierarchical feature regression (HFR) is a novel graph-based regularized regression estimator, which mobilizes insights from the domains of machine learning and graph theory to estimate robust parameters for a linear regression. The estimator constructs a supervised feature graph that decomposes parameters along its edges, adjusting first for common variation and successively incorporating idiosyncratic patterns into the fitting process. The graph structure has the effect of shrinking parameters towards group targets, where the extent of shrinkage is governed by a hyperparameter, and group compositions as well as shrinkage targets are determined endogenously. The method offers rich resources for the visual exploration of the latent effect structure in the data, and demonstrates good predictive accuracy and versatility when compared to a panel of commonly used regularization techniques across a range of empirical and simulated regression tasks.

1. Introduction

In this paper, I propose a new solution to the old problem of obtaining robust parameter estimates in a high-dimensional regression with nonorthogonal predictors. I decompose ordinary least squares (OLS) regression estimates along a supervised hierarchical graph, then optimally shrink the edges of the graph to achieve a group-wise regularization of the parameter space. The resulting estimator has several useful properties: (i) It solves the problem of group shrinkage in an elegant and efficient manner, where the composition of parameter groups as well as group shrinkage targets are determined endogenously; (ii) The estimator offers intuitive tools for the visual inspection of the model effects structure; (iii) It exhibits significant versatility, performing well (in terms of prediction accuracy) both in sparse, as well as dense regression settings; Finally, (iv) the estimator encodes the prior expectation of a world governed by hierarchical processes, making it uniquely suitable for several empirical applications, particularly in the domains of economics and finance.

With increasing availability of data, regularized regressions have steadily grown in importance in many fields, and underpin developments in domains as seemingly disparate as bioinformatics, finance or deep learning. Economic applications in particular are often characterized by high-dimensional, multicollinear data sets, and regularized machine learning algorithms are well established as computationally efficient means of obtaining accurate parameter estimates when the number

E-mail address: johann.pfitzinger@gmail.com

[1] This paper is a revised version of a chapter of my doctoral thesis submitted at Goethe University Frankfurt. I thank my supervisor Uwe Hassler for his advice.

https://doi.org/10.1016/j.ecosta.2024.01.003

of predictors relative to observations is high. The hierarchical feature regression (HFR) contributes to this body of knowledge, combining elements of graph theory and machine learning to inform a novel group shrinkage estimator.

The HFR constructs a parsimonious information graph, using a supervised hierarchical clustering algorithm that groups predictors based on the similarity of their explanatory content with respect to a dependent variable. The information graph is translated into a parameter hierarchy, consisting of several chains of coefficients (edges in the graph) that capture increasingly nuanced signal. The coefficient chains adjust first for shared variation, with each lower element introducing a further degree of idiosyncrasy. By shrinking the chain of coefficients, the HFR achieves group shrinkage — removing idiosyncratic information from the fitting process and giving a higher weight to shared effect patterns.

An economic case study highlights how the structure introduced by the hierarchical graph can be exploited to garner insights into latent effect dynamics in the fitted model, with rich resources for visual exploration. Furthermore, the HFR exhibits robust predictive accuracy, comparing favorably to a panel of benchmark regularized regression techniques. The results also indicate a high degree of versatility in the simulated setting, with good performance across different types of regression settings (e.g. sparse, latent factors, grouped). This flexibility can be a key advantage over related methods that tend to be well-suited to specialized types of data generating processes.

The remainder of this paper is structured as follows: Section 2 introduces important literature relating to the field of regularized regression. The HFR is developed in Section 3, while Sections 4 and 5 explore its performance both in empirical and simulated settings. Finally, Section 6 concludes the paper.

2. Literature review

Nobel prize laureate Herbert Simon posits that complex systems tend to evolve in a hierarchic manner and, as a result, encompass hierarchical structures (Simon, 1962). This proposition is supported by an understanding of highly integrated markets and economies driven by deeper global undercurrents — e.g. global business cycles (Diebold and Yilmaz, 2015; Kose et al., 2003) or global financial cycles (Rey, 2015) —, and is reflected in the popularity of latent variable methods (e.g. dynamic factor models for macroeconomic analysis) and, increasingly, deep learning methods for nonlinear prediction tasks.[2]

The HFR utilizes empirical data hierarchies with the objective of achieving an optimal group shrinkage that captures the hierarchical nature of the data generating process and, in turn, attains more robust out-of-sample performance. It is therefore located squarely within the regularization literature. A plethora of approaches to parameter regularization have been developed in this domain. Penalized regressions — termed “Lasso and friends” in Varian (2014) — receive some attention in this paper as natural benchmarks for the HFR. The approaches introduce a constraint on the parameter norm, by adding a penalty function $P_\lambda(\boldsymbol{\beta})$ to the least squares loss of a regression of y on $\mathbf{x}$:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[N^{-1}(y - \mathbf{x}\boldsymbol{\beta})^\top (y - \mathbf{x}\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}) \right]. \tag{2.1}$$

Here $\hat{\boldsymbol{\beta}}$ is a vector of parameter estimates and N is the sample size. The penalty function depends on a hyperparameter λ governing the weight given to the penalty, and typically takes the form $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_i |\beta_i|^q$, where $q = 1$ is a lasso and $q = 2$ is a ridge regression. Important contributions to this literature include James and Stein (1961), Hoerl (1962), Hoerl and Kennard (1970), Tibshirani (1996) and Efron et al. (2004), as well as multiple variants, including Zou and Hastie (2005), Zou (2006) and Zou and Zhang (2009). An introductory overview is found in Friedman et al. (2001).

Penalized regressions — particularly those based on the ℓ_1-norm ($q = 1$) — have been extended to permit group shrinkage in the case when prior knowledge of predictor (or response) clusters is available (Bach et al., 2012; Huang and Zhang, 2010; Tibshirani et al., 2005; Turlach et al., 2005). Group shrinkage can take various forms. Letting G be the number of groups, the most common approach encourages structured sparsity, (i) by shrinking entire groups of coefficients to zero — known as the group lasso with $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{g=1}^{G} ||\boldsymbol{\beta}_g||_2^1$ (Roth and Fischer, 2008; Yuan and Lin, 2006); or, (ii) its converse the exclusive lasso with $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{g=1}^{G} ||\boldsymbol{\beta}_g||_1^2$, which induces within-group sparsity (Campbell and Allen, 2017; Zhou et al., 2010).[3] Extensions to the group lasso can accommodate more complex prior knowledge, such as shrinkage across as well as within groups (Simon et al., 2013; Szafranski et al., 2007), overlapping groups (Jacob et al., 2009; Zhao et al., 2009) or predictor graphs (Huang et al., 2011; Kim and Xing, 2012; Li and Li, 2008; 2010; Shen et al., 2012).

An alternative conception of group shrinkage aims to shrink clusters of coefficients towards a common target. The most straightforward method is to define a clustered ℓ_2-norm of the form

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{g=1}^{G} \frac{1}{|C_g|} \sum_{i,j \in C_g} ||x_i\beta_i - x_j\beta_j||^2, \tag{2.2}$$

where there are G groups and C_g indexes predictors in group g (Witten et al., 2014). Eq. 2.2 can be altered to include a sparsity-inducing term (the clustered elastic net in Witten et al. (2014)), or to use alternative grouping norms, such as in

[2] Deep neural networks, for instance, have been described as nonlinear hierarchical feature methods (Mishra and Gupta, 2017).

[3] This notation uses the ℓ_1-/ℓ_2-norm shorthand employed, for instance, in Qiu et al. (2021) to denote an ℓ_1-norm (sparsity) across G groups and an ℓ_2-norm (no sparsity) within groups in the case of the group lasso, and the opposite in the case of the exclusive lasso.

Table 2.1
Examples of admissible coefficient estimates for different group shrinkage estimators. $\hat{\beta}_1$ and $\hat{\beta}_2$ are coefficient estimates for features belonging to group *A*, and $\hat{\beta}_3$ and $\hat{\beta}_4$ for features belonging to group *B*.

	Group *A*		Group *B*	
Group lasso (Yuan and Lin, 2006)	$\hat{\beta}_1 = 0$	$\hat{\beta}_2 = 0$	$\hat{\beta}_3 \neq 0$	$\hat{\beta}_4 \neq 0$
Exclusive lasso (Zhou *et al.*, 2010)	$\hat{\beta}_1 = 0$	$\hat{\beta}_2 \neq 0$	$\hat{\beta}_3 = 0$	$\hat{\beta}_4 \neq 0$
Fused lasso (Tibshirani *et al.*, 2005)	$\hat{\beta}_1 \approx \hat{\beta}_2$	$\hat{\beta}_2 \approx \hat{\beta}_1$	$\hat{\beta}_3 = 0$	$\hat{\beta}_4 = 0$
Hierarchical penalty (Szafranski *et al.*, 2007)	$\hat{\beta}_1 = 0$	$\hat{\beta}_2 = 0$	$\hat{\beta}_3 = 0$	$\hat{\beta}_4 \neq 0$
Graph Laplacian (Huang *et al.*, 2011)	$\hat{\beta}_1 \approx \hat{\beta}_2$	$\hat{\beta}_2 \approx \hat{\beta}_1$	$\hat{\beta}_3 = 0$	$\hat{\beta}_4 = 0$
Clustered ℓ_2-norm (Witten *et al.*, 2014)	$\hat{\beta}_1 \approx \hat{\beta}_2$	$\hat{\beta}_2 \approx \hat{\beta}_1$	$\hat{\beta}_3 \approx \hat{\beta}_4$	$\hat{\beta}_4 \approx \hat{\beta}_3$
OSCAR (Bondell and Reich, 2008)	$\hat{\beta}_1 = \hat{\beta}_2$	$\hat{\beta}_2 = \hat{\beta}_1$	$\hat{\beta}_3 = 0$	$\hat{\beta}_4 \neq 0$
Hierarchical feature regression	$\hat{\beta}_1 \approx \hat{\beta}_2$	$\hat{\beta}_2 \approx \hat{\beta}_1$	$\hat{\beta}_3 \approx \hat{\beta}_4$	$\hat{\beta}_4 \approx \hat{\beta}_3$

the OSCAR[4] and related methods (Bondell and Reich, 2008; Sharma et al., 2013; Zeng and Figueiredo, 2013). Graph-based estimators including the graph Laplacian (Huang et al., 2011; Li and Li, 2008; 2010) or the fused lasso (Daye and Jeng, 2009; Tibshirani et al., 2005) represent hybrids between structured sparsity and the group target shrinkage of Eq. 2.2. A sparsity inducing ℓ_1-penalty is augmented by a smoothness term over coefficients sharing an edge in a graph, leading to a penalty similar to Eq. 2.3, where E is an edge set. The smoothness penalty uses, for instance, correlation-based graphs (Daye and Jeng, 2009) or a Laplacian matrix (Huang et al., 2011; Li and Li, 2010), and — as in the case of group target shrinkage — results in group-conformity (i.e. estimates for coefficients within a group are similar).

$$P_{\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \lambda_1 \sum_i |\beta_i| + \lambda_2 \sum_{i,j \in E} |\beta_i - \beta_j|. \tag{2.3}$$

Table 2.1 displays admissible coefficient values under different types of group shrinkage, including both structured sparsity and group target shrinkage estimators, illustrating the nature of regularization in each.

The reliance on prior knowledge introduces a risk of misspecification to group shrinkage methods and limits their usefulness when no obvious *ex ante* grouping is available. While methods have been proposed to estimate an appropriate grouping using clustering algorithms in conjunction with the group lasso (Bühlmann et al., 2013; Grimonprez et al., 2022) or in conjunction with a clustered ℓ_1/ℓ_2-norm (Witten et al., 2014), Section 5.1 shows that misspecification remains a concern.

The HFR proposes a solution to this issue by using a data-driven approach to estimate a supervised hierarchy, and subsequently performing endogenous selection of the appropriate group constellations (levels in the hierarchy) to avoid misspecification. Eq. 2.4 illustrates the concept by generalizing Eq. 2.2 to include L level-specific penalties weighted by θ_ℓ, with $G^{(\ell)}$ and $C_g^{(\ell)}$ defined for the ℓth level:

$$P_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}) = \lambda \left(\sum_{\ell=1}^{L} \theta_\ell \sum_{g=1}^{G^{(\ell)}} \frac{1}{|C_g^{(\ell)}|} \sum_{i,j \in C_g^{(\ell)}} ||x_i\beta_i - x_j\beta_j||^2 \right). \tag{2.4}$$

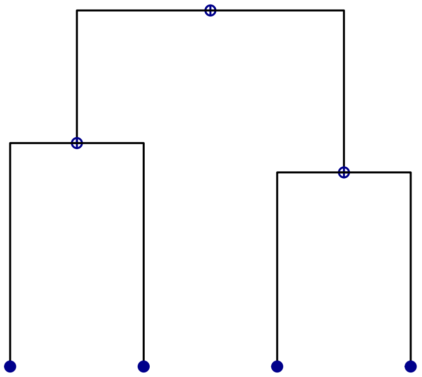
The increased complexity of Eq. 2.4 makes estimation virtually unfeasible. However, a key insight of the proposed HFR is to replace the level-specific soft constraints with hard constraints, which reduces the problem to a quadratic program and yields an elegant graphical interpretation of the estimator, as well as an intuitive hyperparameter in the form of a complexity constraint, with θ_ℓ becoming the ℓth additive contribution to the estimator's effective degrees of freedom.

The type of complexity constraint imposed by the HFR is closely related to a second broad class of regularization techniques: latent variable regressions. Examples include the principal components regression (PCR) described in Friedman et al. (2001), the partial least squares regression (PLSR) developed by Wold in the 1960s and 70s (see Wold (2001) and Martens (2001)), or — in the econometric setting — the dynamic factor model surveyed in Stock and Watson (2016a) and Stock and Watson (2016b).

These methods reduce the dimensionality of the predictor set by removing low variance components in the case of principal components based methods, or components with a low response correlation in the case of PLSR (Jolliffe, 2002). Unlike penalized regressions, latent variable regressions are mostly unsupervised in their construction of latent factors. Some exceptions exist, for instance the aforementioned PLSR, or Bair et al. (2006), who introduce a (semi-)supervised PCR.

In contrast to traditional factor methods, the HFR proposes a hierarchical transformation of the predictor space. The concept of feature hierarchies has been applied in the machine learning domain to visual and text classification tasks, where general features (e.g. objects, phrases) are learned first, with subsequent fine-tuning for lower level representations (e.g. pixels, words) (Epshtein and Uliman, 2005; Girshick et al., 2014). Porting this concept to the linear regression setting, the HFR decomposes the data generating process (DGP) into a signal graph, estimating parameters for general (shared) signal patterns separately from the idiosyncratic contribution of each individual predictor.

[4] The Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) uses a penalty function of the form $P_{\lambda,c}(\boldsymbol{\beta}) = \sum_i |\beta_i| + c\sum_{j<k} \max\{|\beta_j|, |\beta_k|\}$.



$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Fig. 3.1. Example of a hierarchy dendrogram (left) and the corresponding summing matrix (right).

To illustrate the implicit hierarchical structure of the penalty, Section 3 begins with a derivation of the HFR as a hierarchical decomposition of a least squares estimator, and then proceeds to discuss the resulting regularized regression equation.

3. The HFR estimator

3.1. Syntax of feature hierarchies

Before introducing the HFR estimator, this section provides a brief overview of the graph theoretical concepts and definitions drawn on in the subsequent discussions.

A hierarchical representation is taken to mean the arrangement of predictors into clusters of two or more, which are merged at nodes to form higher levels. The predictors are the leaf nodes (i.e. they represent the lowest nodes in the hierarchy), while nodes at higher levels are called internal nodes. The process of merging is repeated at each level until all predictors are contained within a single cluster called the root node. The node directly above any node is typically referred to as the parent node, while the nodes below are the children. Adjacent nodes that share a single parent are siblings. The chain of preceding parent nodes for any node is its branch.

The nested grouping structure of a hierarchy is captured graphically in a dendrogram, or mathematically in a summing matrix (Hyndman et al., 2011). Fig. 3.1 portrays a simple hierarchy dendrogram of the illustration introduced in Section 3.2. There are $K = 4$ predictors (leaf nodes), and two subsets grouping two predictors each. The root node completes the dendrogram.

The corresponding hierarchy summing matrix $\mathbf{S} = \{s_{ij}\}$ (right panel, Fig. 3.1) consists of $D \times K$ dimensions, where $D = 7$ is the total number of nodes and $K = 4$ is the number of predictors. Any entry s_{ij} in $\mathbf{S}$ is defined such that

$$s_{ij} = \begin{cases} 1 & \text{when } j \text{ is a descendant of } i \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbf{S}$ is invariant to the ordering of rows (i.e. child and parent nodes do not have to be arranged in any particular order). However, to simplify the discussion it is presented in a top-down order throughout this paper, starting with the root node and ending with the leaf nodes.

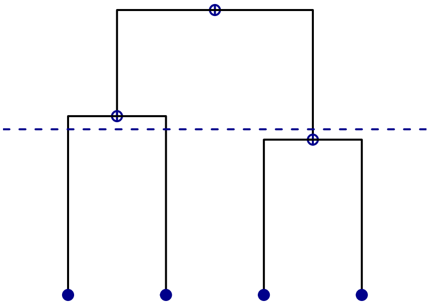
Hierarchies can be cut along the y-axis of the dendrogram by drawing a horizontal line at any height of Fig. 3.1.[5] The nodes directly beneath the cut describe a level. In the discussions that follow, an arbitrary level is denoted ℓ, and L is the total number of levels. Fig. 3.2 shows a cut in the dendrogram and the summing matrix associated with that level, where the summing matrix is the submatrix of $\mathbf{S}$ containing the rows corresponding to nodes directly beneath the given cut.

A predictor hierarchy conveys information about the interrelatedness of predictors, grouping similar predictors closely together. Sections 3.2 to 3.5 illustrate how the HFR exploits this information to achieve a type of group shrinkage under the assumption of a given optimal hierarchy $\mathbf{S}$, while Section 3.6 introduces an algorithm to estimate $\mathbf{S}$.

3.2. A framework for group shrinkage

The HFR assumes that similar predictors are imperfect measures of one or more underlying latent concepts, which drive the true model, as captured in Assumptions 1 & 2:

[5] It is assumed that no two nodes can be located at the exact same height in the dendrogram (i.e. it is always possible to separate nodes into those above and those below a horizontal line). This holds for the hierarchical clustering algorithms employed in this paper.



$$\mathbf{S}_\ell = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Fig. 3.2. Example of a cut hierarchy dendrogram (left) and the corresponding level-specific summing matrix (right).

Assumption 1. Predictors can be partitioned into groups that represent common latent factors, with moderate to high levels of correlation among predictors in a group. Individual predictors can encompass broad latent factors (large clusters), alongside idiosyncratic factors (smaller local clusters), resulting in a hierarchical nested grouping structure.

Assumption 2. Predictors within a group have a similar effect on the response, reflected in similar coefficient values.

Shrinkage towards group targets for clusters of related variables is expected — under Assumptions 1 & 2 — to lead to a more efficient bias-variance trade-off than can be attained by methods that discard hierarchical information.[6] The HFR achieves this type of shrinkage by arranging predictors in a hierarchical graph, with coefficients on predictors whose paths merge experiencing shrinkage towards a common target. The higher in the hierarchy the merge is located, the stronger the shrinkage.

The proposed estimator is introduced in this section using a simple example, and following two steps: First, a decomposition of the OLS estimator into a sequence of node-specific estimates in a hierarchical graph is proposed. Second, shrinkage is introduced to the levels of the graph, resulting in the HFR estimator.

Take again the setting described above with $K = 4$ standardized predictors, $\mathbf{x} = \{\mathbf{x}_i\}_{i=1,\ldots,N} \in \mathbb{R}_K$, where $\mathbf{x}_i$ is a row-vector of length K associated with the ith observation. The predictors are clustered into one, two, three and four groups, as captured in the hierarchy in Fig. 3.1, assumed to represent an optimal graph. The matrix $\mathbf{S}$ stacks the individual levels, and can be divided into sub-matrices, denoted $\mathbf{S}_\ell$, that describe the individual levels within the feature hierarchy.

For the four levels in the example, with $\ell = 1, 2, 3, 4$, the sub-matrices are given by

$$\mathbf{S}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} \quad \mathbf{S}_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \mathbf{S}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \mathbf{S}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Here the lowest level ($\mathbf{S}_4$) is an identity matrix containing the leaf nodes.

The level-specific hierarchical features are now defined as $\mathbf{z}_\ell = \mathbf{x}\mathbf{S}_\ell^\top$, and the complete hierarchical feature set is given by $\mathbf{z} = \mathbf{x}\mathbf{S}^\top = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \mathbf{z}_3 & \mathbf{z}_4 \end{bmatrix}$. The hierarchical features, $\mathbf{z}$, represent factor estimates of the common variance contained in the child features (i.e. the features associated with child nodes). Under the assumption that the covariance between the predictors' idiosyncratic components is low (such that their mean converges to zero), the sum of the predictors represents an estimate of the common component that is consistent up to a constant scale. See Stock and Watson (2016b) for a discussion of the role of feature averaging in factor estimation.

Using the level-specific factor estimates, define

$$\mathbf{Q}_{ij} = \mathbf{z}_i^\top \mathbf{z}_j \quad \text{and} \quad \tilde{\mathbf{Q}}_{\ell y} = \mathbf{z}_\ell^\top \mathbf{M}_{\ell-1} y, \quad i, j, \ell \in \{1, 2, 3, 4\},$$

with the regression response variable $y = \{y_i\}_{i=1,\ldots,N} \in \mathbb{R}$. Here $\mathbf{M}_\ell$ is the residual maker matrix, with $\mathbf{M}_\ell = \mathbf{I}_N - \mathbf{P}_\ell = \mathbf{I}_N - \mathbf{z}_\ell \mathbf{Q}_{\ell\ell}^{-1} \mathbf{z}_\ell^\top$, and $\mathbf{M}_0 = \mathbf{I}_N$. Furthermore, $\mathbf{I}_N$ is an $N \times N$ dimensional identity matrix. The role of $\mathbf{M}_{\ell-1}$ is to partial out the effect of each node's branch from $\tilde{\mathbf{Q}}_{\ell y}$, resulting in a regression that updates parameter estimates using only the new information introduced at each level. Note that in a nested hierarchical graph where each level contains strictly more information than the preceding levels, it holds that

$$\mathbf{M}_{\ell-1} \equiv \prod_{i=1}^{\ell} \mathbf{M}_{\ell-i}. \tag{3.1}$$

[6] The HFR estimation produces an intuitive visual check for the validity of Assumption 1 in the form of level-wise contributions to total explained variance, as described in Section 3.7.

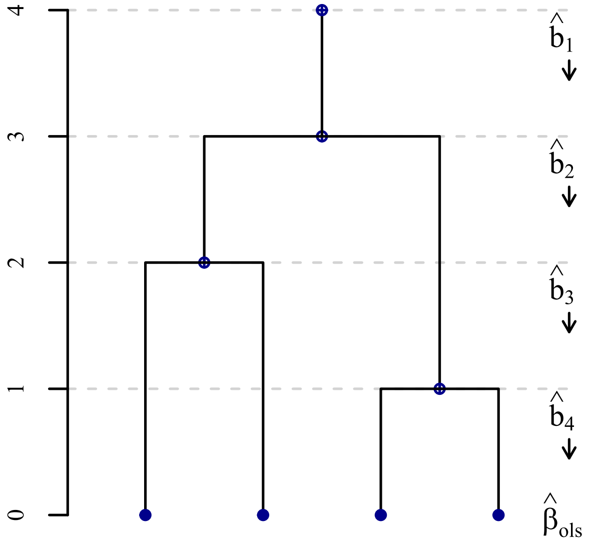


Fig. 3.3. Dendrogram of the level-wise decomposition of the OLS estimator.

Thus, the information of the entire branch can be partialled out using only $\mathbf{M}_{\ell-1}$.

Now, with $\hat{\boldsymbol{\beta}}_{\text{ols}}$ denoting OLS estimates for a regression of y on $\mathbf{x}$, a top-down hierarchical decomposition of the OLS estimator for our problem is given by

$$\hat{\boldsymbol{\beta}}_{\text{ols}} = \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2 + \hat{\mathbf{b}}_3 + \hat{\mathbf{b}}_4, \tag{3.2}$$

where $\hat{\mathbf{b}}_\ell$ are level-specific estimates that account for the new variation introduced at level ℓ. The level-specific estimates are defined simply as the least squares estimates for $\mathbf{z}_\ell$ conditional on the path of each node:

$$\hat{\mathbf{b}}_\ell = \mathbf{S}_\ell^\top \mathbf{Q}_{\ell\ell}^{-1} \tilde{\mathbf{Q}}_{\ell y}. \tag{3.3}$$

Proposition 1 stacks the above decomposition, and shows that the resulting estimates are algebraically equivalent to OLS estimates:

Proposition 1. *Consider a regression decomposition for a hierarchy with L levels described by summing matrix $\mathbf{S} = \begin{bmatrix}\mathbf{S}_1^\top & \cdots & \mathbf{S}_L^\top\end{bmatrix}^\top$, hierarchical features $\mathbf{z}$ defined as above, as well as $\mathbf{Q}_{ij} = \mathbf{z}_i^\top \mathbf{z}_j$, $\mathbf{Q}_{zy} = \mathbf{z}^\top y$ and*

$$\mathbf{Q}_{zz} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Q}_{L1} & \mathbf{Q}_{L2} & \cdots & \mathbf{Q}_{LL} \end{bmatrix}.$$

Now the coefficient estimates $\hat{\boldsymbol{\beta}} = \mathbf{S}^\top \mathbf{Q}_{zz}^{-1} \mathbf{Q}_{zy}$ represent optimal least squares estimates of the linear slope coefficients $\boldsymbol{\beta}$ of a regression of y on $\mathbf{x}$.

The proof of Proposition 1 is given in Appendix A. Note that $\mathbf{Q}_{zz}$ can be written as $\mathbf{Q}_{zz} = (\mathbf{z}^\top \mathbf{z}) \odot \mathbf{H}$, where $\odot$ is the element-wise multiplication operator, and $\mathbf{H}$ is a matrix of ones with the block-wise upper triangle set to zero:

$$\mathbf{H} = \begin{bmatrix} \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{1} & \cdots & \mathbf{1} \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{0} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}.$$

The matrix $\mathbf{H}$ eliminates bottom-up conditional effects from the precision matrix, which are represented by the upper block-triangular entries. Conversely, the lower block-triangular entries represent conditional effects flowing down the hierarchy from the root node towards the leaf nodes (i.e. top-down effects). Setting the upper block-triangular entries to zero therefore partials out each node's branch and has the equivalent effect as $\mathbf{M}_{\ell-1}$ in $\hat{\mathbf{b}}_\ell$.

Fig. 3.3 plots a dendrogram of the decomposition used in the example, expanding the root node so that each level is represented by a band of unit width. As shown in the next subsection, the width of each level-specific band will come

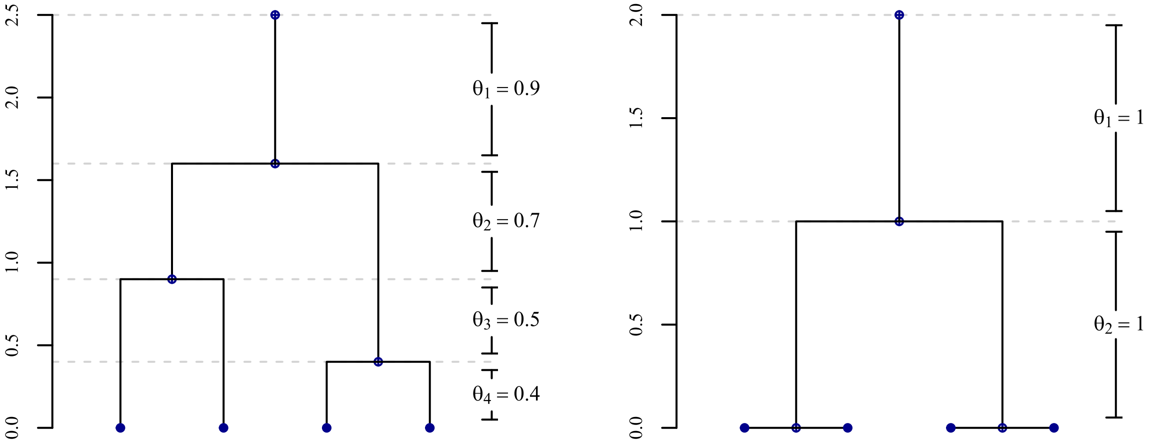


Fig. 3.4. Dendrogram of the level-wise decomposition of the OLS estimator with shrinkage represented by the distance between levels.

to represent the proportion to which information introduced at that level is incorporated into the HFR estimates. Each $\hat{\mathbf{b}}_\ell$ adjusts the coefficients based on the new idiosyncratic variation at ℓ, with a single cluster at $\hat{\mathbf{b}}_1$, two clusters at $\hat{\mathbf{b}}_2$, etc., until at the final level ($\ell = 4$ in the example) all explainable variation is accounted for.

3.3. Adding shrinkage coefficients

While this decomposition seems trivial at first glance, it can be used as the basis for a regularized regression by shrinking the extent to which levels are permitted to adjust for new variation. This results in estimates that are biased towards higher-level representations in the form of group targets for clusters of predictors.

In the simplest form, one could add shrinkage coefficients to Eq. 3.2, such that

$$\hat{\boldsymbol{\beta}}_{\text{hfr}} = \sum_{\ell=1}^{L} \theta_\ell \hat{\mathbf{b}}_\ell, \quad \text{with } 0 \le \theta_\ell \le \theta_{\ell-1} \text{ and } 0 \le \theta_1 \le 1, \tag{3.4}$$

where θ_ℓ is the ℓth shrinkage coefficient.

The shrinkage coefficients permit a variety of regularized solutions for $\hat{\boldsymbol{\beta}}_{\text{hfr}}$. Exact group constraints (i.e. all coefficients in a group are equal, $\hat{\boldsymbol{\beta}}_{\text{hfr}} = \hat{\mathbf{b}}_\ell$) result when $\theta_i = 1 \ \forall \ i \le \ell$ and $\theta_i = 0 \ \forall \ i > \ell$. Shrinkage towards a common group target (soft constraint) results when any $0 < \theta_\ell < 1$. The monotonicity constraint on θ_ℓ ensures that — given that the hierarchy represents a nested information set — information that is removed at one level is not subsequently reintroduced at a lower level.

Fig. 3.4 plots two shrunken dendrograms with the degree of shrinkage represented by the vertical distance between two levels and given by θ_ℓ. The left panel of Fig. 3.4 represents moderate shrinkage towards group targets, while the right panel removes two entire levels and results in an exact group constraint with $\hat{\beta}_1 = \hat{\beta}_2$ and $\hat{\beta}_3 = \hat{\beta}_4$.

In the stacked form of Proposition 1, Eq. 3.4 introduces a shrinkage matrix, such that

$$\hat{\boldsymbol{\beta}}_{\text{hfr}} = \mathbf{S}^\top \boldsymbol{\Theta} \mathbf{Q}_{zz}^{-1} \mathbf{Q}_{zy}. \tag{3.5}$$

Here $\boldsymbol{\Theta}$ is a $D \times D$ diagonal matrix governing the extent of shrinkage, with

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Theta}_L \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Theta}_\ell = \theta_\ell \mathbf{I}_\ell.$$

$\boldsymbol{\Theta}_\ell$ is a diagonal submatrix, with the dimensions of the identity matrix $\mathbf{I}_\ell$ given by the number of nodes in the ℓth level.

Eq. 3.5 is the HFR estimator under the assumption that the hierarchy ($\mathbf{S}$) as well as the extent of shrinkage ($\boldsymbol{\Theta}$) are given. Since the framework permits the exclusion of entire levels from the regression (by setting $\theta_\ell = 0$), it can be used as a tool to select a parsimonious hierarchy based on a (potentially large) set of input levels. As shown in subsequent sections, this property will be useful for the estimation of $\mathbf{S}$.

The framework described by Eq. 3.5 can become arbitrarily complex, including a large number of levels that permit a high degree of nuance with respect to the nature and strength of regularization. At its core, however, it remains a decomposition of each parameter into a chain of parameters that captures successively more idiosyncratic signal, and which is subsequently regularized, resulting in overall shrinkage towards a more general and less idiosyncratic representation of the data generating process.

A key ingredient for this form of group shrinkage is determining the optimal extent of shrinkage for each hierarchical level (i.e. for the elements of the parameter chain). Section 3.4 discusses an appropriate loss function that can be used to obtain optimal shrinkage coefficients.

3.4. Optimal shrinkage

The extent of shrinkage is governed entirely by the $L \times 1$ vector of level-specific shrinkage coefficients $\boldsymbol{\theta} = \begin{pmatrix}\theta_1 & \cdots & \theta_L\end{pmatrix}$, consisting of the values that make up the diagonal of $\boldsymbol{\Theta}$. When $\boldsymbol{\theta} = \mathbf{1}$ there is no shrinkage, leading to the OLS solution as shown in Proposition 1. If any $\theta_\ell < 1$, the parameters associated with level ℓ are regularized, where $\theta_\ell = 0$ constitutes maximum shrinkage. Note that when $\theta_1 < 1$, the entire parameter-norm is shrunken.

When $\mathbf{S}$ includes the maximum possible number of levels, with $L = K$ so that each level comprises exactly one more cluster than the preceding level, the shrinkage vector $\boldsymbol{\theta}$ has the useful property that its sum (i.e. the sum of all level-specific adjustments to new variation) is equal to the effective model size, as captured by the effective degrees of freedom (ν_{eff}):

Proposition 2. *With an HFR projection matrix given by $\mathbf{P}_{\text{hfr}} = \mathbf{z}\boldsymbol{\Theta}\mathbf{Q}_{zz}^{-1}\mathbf{z}^\top$, and the effective model degrees of freedom defined in the usual manner using the trace of the projection matrix, $\nu_{\text{eff}} = \text{tr}(\mathbf{P}_{\text{hfr}})$, it holds that*

$$\nu_{\text{eff}} = \sum_{\ell=1}^{L} \theta_\ell,$$

when $L = K$ distinct levels are included in the hierarchy described by $\mathbf{S}$.

The proof of Proposition 2 is given in Appendix B.

With this definition in hand, an information theoretically motivated approach to the determination of an optimal shrinkage vector, $\boldsymbol{\theta}^*$, is to impose a constraint on the effective model size. Defining a hyperparameter, κ, that represents the effective model size (normalized to a value between 0 and 1), the optimal shrinkage vector is the solution that maximizes fit subject to the constraint

$$\sum_{\ell=1}^{L} \theta_\ell = \kappa K. \tag{3.6}$$

When $\kappa = 1$, the solution is unconstrained, with $\nu_{\text{eff}} = K$ and $\hat{\boldsymbol{\beta}}_{\text{hfr}} = \hat{\boldsymbol{\beta}}_{\text{ols}}$. Any $\kappa < 1$ results in a biased solution that maximizes model fit subject to a constraint on ν_{eff}. Expressing the optimization in terms of the HFR loss, the optimal extent of shrinkage conditional on hyperparameter κ, is given by

$$\begin{aligned} \boldsymbol{\theta}_\kappa^* = \arg\min_{\boldsymbol{\theta}} & \left[N^{-1}(\mathbf{x}\hat{\boldsymbol{\beta}}_{\text{hfr}} - y)^\top(\mathbf{x}\hat{\boldsymbol{\beta}}_{\text{hfr}} - y)\right] \\ \text{s.t.} \quad & 0 \le \theta_\ell \le \theta_{\ell-1} \ \forall\ \ell > 1, \\ & 0 \le \theta_1 \le 1 \quad \text{and} \\ & \sum_{\ell=1}^{L} \theta_\ell = \kappa K. \end{aligned} \tag{3.7}$$

Eq. 3.7 trades off goodness-of-fit against parsimony, where the hyperparameter κ tilts the global trade-off towards goodness-of-fit as $\kappa \to 1$, or parsimony as $\kappa \to 0$. Fig. 3.5 visualizes the effect using the dendrogram of the example problem with $L = 4$ levels. The total height of the dendrogram is now exactly equal to the effective model size (left panel: $\kappa = 1.00$, right panel: $\kappa = 0.75$). The definition of κ ensures that the hyperparameter represents the overall size of the optimal HFR graph as a percentage of K, with a shallower hierarchy as $\kappa \to 0$.

The following section demonstrates how the optimal shrinkage vector $\boldsymbol{\theta}_\kappa^*$ can be obtained efficiently for any given value of κ using quadratic programming to solve Eq. 3.7.

3.5. Solution algorithm

The HFR estimates in Eq. 3.5 can be restated as the dot product of level-specific estimates and a transformed shrinkage vector, such that

$$\hat{\boldsymbol{\beta}}_{\text{hfr}} = \hat{\boldsymbol{\mathcal{B}}}\boldsymbol{\phi}. \tag{3.8}$$

Here $\hat{\boldsymbol{\mathcal{B}}}$ stacks unconditional level-specific estimates (unconditional with respect to preceding levels in the hierarchy), such that with $\hat{\boldsymbol{\mathcal{B}}} = \begin{bmatrix}\hat{\mathbf{w}}_1 & \cdots & \hat{\mathbf{w}}_L\end{bmatrix}$,

$$\hat{\mathbf{w}}_\ell = \mathbf{S}_\ell^\top\left(\mathbf{z}_\ell^\top\mathbf{z}_\ell\right)^{-1}\mathbf{z}_\ell^\top y. \tag{3.9}$$

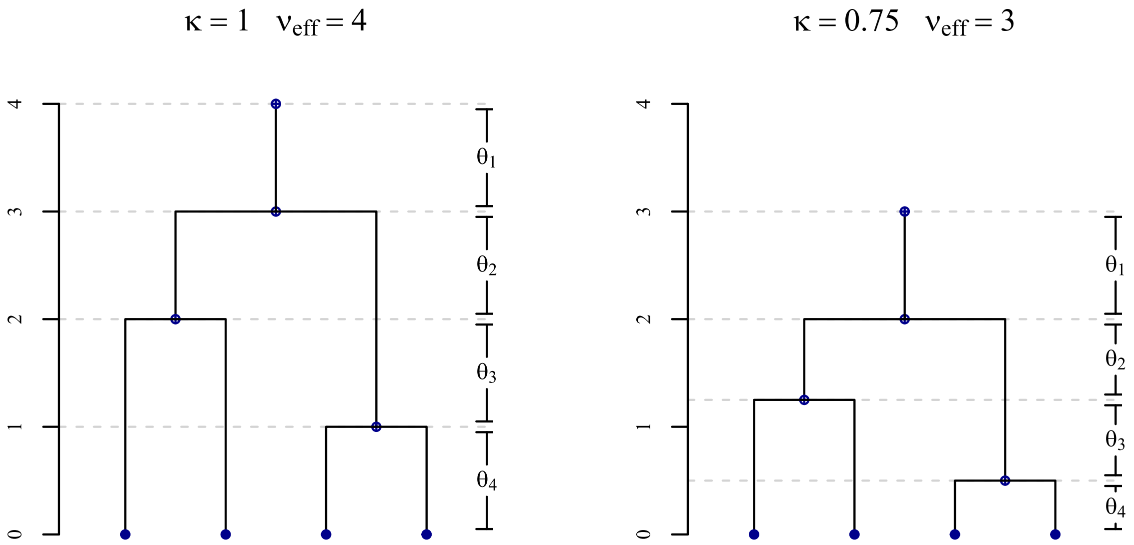


Fig. 3.5. Dendrogram of the level-wise decomposition of the OLS estimator with $L = K$ levels, and $\kappa = 1$ (left panel) and $\kappa = 0.75$ (right panel).

Note that $\hat{\mathbf{w}}_\ell$ is an unconditional counterpart to $\hat{\mathbf{b}}_\ell$, where the effect of each node's branch has not been partialled out. Furthermore, $\boldsymbol{\phi}$ is a transformation of $\boldsymbol{\theta}$ that satisfies the equality $\boldsymbol{\theta} = \boldsymbol{\omega}^\top \boldsymbol{\phi}$, where $\boldsymbol{\omega}$ is a lower triangular matrix, resulting in

$$\boldsymbol{\phi} = \begin{cases} \theta_\ell - \theta_{\ell+1} & \text{when } \ell < L \\ \theta_\ell & \text{otherwise.} \end{cases}$$

The derivation of Eq. 3.8 is given in Appendix C, and follows directly from the introduction of shrinkage weights to the calculations in Appendix A. By reformulating the problem in an unconditional manner, $\hat{\mathbf{w}}_\ell$ can be computed in parallel for each level, and the optimization of $\boldsymbol{\theta}$ can be split into two consecutive steps: (i) estimating level-specific regressions ($\hat{\mathcal{B}}$), and (ii) constructing the optimal shrinkage hierarchy by optimizing $\boldsymbol{\phi}$.

Eq. 3.8 resembles a model-averaging estimator, where the models $\hat{\mathcal{B}}$ are averaged by the weights $\boldsymbol{\phi}$. Mallows model averaging (MMA), for instance, represents a close mathematical pendant, where the weighting vector is obtained by minimizing the Mallows information criterion (Hansen, 2007; Mallows, 1973). The optimal shrinkage problem of the HFR can correspondingly be thought of as the minimization of a custom information criterion (given the constraint penalty on ν_{eff} in Eq. 3.7) to determine the optimal vector $\boldsymbol{\theta}^*_\kappa$. Importantly, the information theoretic model-averaging problem is quadratic in its weights, and can be solved analytically using quadratic programming algorithms.

Following this reasoning, the level-specific coefficients in $\hat{\mathcal{B}}$ are used to reformulate Eq. 3.7, such that with $\hat{\mathbf{y}} = \mathbf{x}\hat{\mathcal{B}}$, $\mathbf{U} = \frac{1}{N}\hat{\mathbf{y}}^\top \hat{\mathbf{y}}$ and $\mathbf{V} = \frac{2}{N}\hat{\mathbf{y}}^\top y$:

$$\begin{aligned} \boldsymbol{\phi}^*_\kappa = \arg\min_{\boldsymbol{\phi}} & \left[\boldsymbol{\phi}^\top \mathbf{U} \boldsymbol{\phi} - \mathbf{V}^\top \boldsymbol{\phi} \right] \\ \text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\phi} \le \mathbf{1} \quad \text{and} \\ & \boldsymbol{\phi}^\top \boldsymbol{\omega} \mathbf{1} = \kappa K. \end{aligned} \tag{3.10}$$

Since the original shrinkage vector can be expressed as $\boldsymbol{\theta} = \boldsymbol{\omega}^\top \boldsymbol{\phi}$, the constraints in Eq. 3.10 are identical to the constraints in Eq. 3.7, and $\boldsymbol{\theta}^*_\kappa = \boldsymbol{\omega}^\top \boldsymbol{\phi}^*_\kappa$. Note that the monotonicity constraint on $\boldsymbol{\theta}$ collapses to a simple weight constraint on $\boldsymbol{\phi}$.

The HFR estimates given by $\hat{\boldsymbol{\beta}}_{\text{hfr}} = \hat{\mathcal{B}} \boldsymbol{\phi}^*_\kappa$ have thus far assumed a given hierarchy, encoded in $\mathbf{S}$. The aim of the HFR is to estimate $\mathbf{S}$ in a supervised manner, which conceptually requires selecting the composition of predictor groups at each level that minimizes Eq. 3.7. This quickly becomes an intractable combinatorial problem. Instead, the following section presents a feasible and computationally efficient algorithm for arriving at a graph estimate based on the similarity of the predictors' explanatory structure in y, using supervised hierarchical clustering.

3.6. Graph estimation

The graph-based decomposition of linear regression parameters introduced in the preceding sections assumes a hierarchical arrangement of predictors into $L = K$ levels that are captured in $\mathbf{S}$. Here $\mathbf{S}$ contains the maximum number of levels possible in a nested hierarchical tree, while $\boldsymbol{\theta}^*_\kappa$ selects a parsimonious hierarchy by reducing the weight of individual levels, or removing levels from the hierarchy entirely. In order to estimate $\mathbf{S}$, I propose a supervised hierarchical clustering algorithm, that merges variables based on the similarity of their explanatory component with respect to y.

A typical approach to (unsupervised) hierarchical clustering constructs a dissimilarity matrix $\mathcal{D}$ that encodes information about the predictor set (e.g. Euclidean distances), and recursively merges the predictors or clusters with the lowest overall cluster distance (Maimon and Rokach, 2010). The aim of supervised hierarchical clustering is to merge those predictors or clusters that maximize the goodness-of-fit of a regression of y on the appropriate cluster features $\mathbf{z}_\ell$ at each ℓ.[7] Two predictors or clusters are deemed similar, if merging them leads to a comparatively small increase in the regression error, or conversely, a comparatively small decline in the goodness-of-fit.

Consider the previous example of a regression of y on four predictors $\mathbf{x}$, with the estimated regression fit given by:

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4. \tag{3.11}$$

Given the definition of $\mathbf{S}$, a merge of any two predictors $i, j \in 1, \ldots, 4$ results in:

$$\hat{y} = \hat{\beta}_{ij}(x_i + x_j) + \mathbf{x}_{-ij}\hat{\boldsymbol{\beta}}_{-ij}, \tag{3.12}$$

where $\mathbf{x}_{-ij}$ contains all remaining predictors.

A merge is therefore akin to the imposition of an equality constraint on the associated coefficients β_i and β_j. This equality constraint is least costly (in terms of goodness-of-fit), when the conditional effect of x_i and x_j on y is similar. That is when

$$r_{x_i,y|\mathbf{x}_{-ij}} \approx r_{x_j,y|\mathbf{x}_{-ij}}. \tag{3.13}$$

Here, $r_{x_i,y|\mathbf{x}_{-ij}}$ is the partial correlation between x_i and y conditional on $\mathbf{x}_{-ij}$.

An intuitively appealing and computationally feasible alternative to the estimation of regression fits for each possible cluster combination (i.e. estimating Eq. 3.12 for all $i \neq j$), is therefore to examine the similarity of the partial correlation coefficients. If the within-cluster variance of partial correlations is small (Eq. 3.13), the reduction in goodness-of-fit can be expected to be low.

Ward (1963) outlines an agglomerative clustering algorithm that merges clusters based on the minimum additional within-cluster variance introduced by the merge. The author shows that the approach can be reduced to a clustering based on the Euclidean distances between the input vectors. The algorithm begins by placing each row in $\mathcal{D}$ into a cluster of its own, and iteratively merges those clusters that result in the minimum increase in overall within-cluster variance. Clusters are merged a total of $K - 1$ times, until all rows in $\mathcal{D}$ are contained in a single cluster, and $L = K$ levels have been formed (Everitt et al., 2011; Kaufman and Rousseeuw, 2005).[8]

Substituting partial correlations for $\mathcal{D}$ results in a supervised hierarchical clustering algorithm, with clusters formed based on the similarity of the predictors' conditional effect on y:

$$\mathcal{D}_y = \{r_{x_i,y|\mathbf{x}_{-ij}}\}_{i,j=1,\ldots,K}, \quad i \neq j, \tag{3.14}$$

with the corresponding distance matrix defined as $|\mathcal{D}_y - \mathcal{D}_y^\top|$.

However, since conditioning on $\mathbf{x}_{-ij}$ is at best imprecise and at worst unfeasible in the high-dimensional setting, $r_{x_i,y|\mathbf{x}_{-ij}}$ needs to be approximated. Two possible approaches include (i) using a shrinkage estimate of $r_{x_i,y|\mathbf{x}_{-ij}}$, denoted $\mathcal{D}_y^{(\mathrm{s})}$ (Ledoit and Wolf, 2003; Schäfer and Strimmer, 2005), or (ii) defining $\mathcal{D}_y$ based on bivariate partial correlations, denoted $\mathcal{D}_y^{(\mathrm{b})}$, such that

$$\mathcal{D}_y^{(\mathrm{b})} = \{r_{x_i,y|x_j}\}_{i,j=1,\ldots,K}, \quad \text{and} \quad r_{x_i,y|x_j} = \frac{r_{y,x_i} - r_{y,x_j} r_{x_i,x_j}}{\sqrt{(1 - r^2_{y,x_j})(1 - r^2_{x_i,x_j})}}, \quad i \neq j. \tag{3.15}$$

Note that $\mathrm{diag}\left(\mathcal{D}_y^{(\mathrm{b})}\right)$ in Eq. 3.15 is undefined so that, letting $\boldsymbol{d}_i$ denote the ith row, $||\boldsymbol{d}_i - \boldsymbol{d}_j||$ measures the distance between $\{r_{x_i,y|x_k}\}_{k \notin i,j}$ and $\{r_{x_j,y|x_k}\}_{k \notin i,j}$ (i.e. the distance between the bivariate partial correlations conditioning on all predictors in $\mathbf{x}_{-ij}$ individually).

The latter approach using $\mathcal{D}_y^{(\mathrm{b})}$ is preferred due to its computational efficiency, and robust performance both in an empirical and simulated setting. Nonetheless, as illustrated in Section 5.3, a more complex DGP may require a fully conditional estimate of $r_{x_i,y|\mathbf{x}_{-ij}}$ to allow meaningful graph estimation.

Regardless of the estimation approach chosen, the matrix $\mathcal{D}_y$ results in a sign-sensitive clustering of parameters (positive and negative coefficients tend to be clustered separately). However, at the highest levels in the hierarchy, clusters will invariably contain effects with mixed signs. To ensure sign-invariance, with shrinkage towards absolute group targets, the summing matrix $\mathbf{S}$ must be adjusted such that

$$\mathbf{S}_i = \mathbf{S}_i^+ \odot \mathrm{sign}(\boldsymbol{\rho}^\top \mathbf{S}_i^+ - \mathbf{1}), \tag{3.16}$$

where $\mathbf{S}_i$ is a row in $\mathbf{S}$, and $\boldsymbol{\rho} = \mathrm{cor}(\mathbf{x})$. The matrix $\mathbf{S}^+$ is the unadjusted (positive-only) summing matrix.

[7] This differs conceptually from a traditional understanding of supervised clustering, where true cluster labels are used to train a model, with the aim of predicting new cluster labels.

[8] The algorithm is implemented, for instance, in the `cluster` package in the statistical computing language R (Maechler et al., 2019; R Core Team, 2018).

Eq. 3.16 ensures that when correlated effects with opposite signs are contained in a single cluster, their coefficient is mirrored and not averaged. If averaged, a common component could be removed (or greatly reduced) in a given hierarchical level, even if it contains a high degree of explanatory variance. The sign adjustment of $\mathbf{S}$ has the desirable effect of retaining important but negatively correlated components.[9]

The combination of Ward (1963) clustering and partial correlations between y and $\mathbf{x}$ produces a supervised hierarchical clustering algorithm that merges clusters based on the within-cluster variance of the partial correlations — analogous to the selection of cluster-splits using a goodness-of-fit criterion in a regression —, but implemented within an efficient agglomerative framework.

3.7. Deterministic terms, standard errors and further issues

The preceding discussions have abstracted from deterministic elements in the regression. Including these is exceedingly simple, and can be achieved by adjusting the level-specific regressions in $\hat{\mathcal{B}}$. Letting $\mathbf{C} = \{\mathbf{C}_i\}_{i=1,\dots,N} \in \mathbb{R}_C$ be a matrix of C deterministic elements (e.g. a vector of ones), with the associated parameter estimates $\hat{\mathbf{c}}$, the level-specific regression becomes:

$$\begin{bmatrix}\hat{\mathbf{c}}_\ell \\ \hat{\mathbf{w}}_\ell\end{bmatrix} = \tilde{\mathbf{S}}_\ell^\top (\tilde{\mathbf{z}}_\ell^\top \tilde{\mathbf{z}}_\ell)^{-1} \tilde{\mathbf{z}}_\ell^\top y, \tag{3.17}$$

where $\tilde{\mathbf{z}}_\ell = \begin{bmatrix}\mathbf{C} & \mathbf{z}_\ell\end{bmatrix}$, and $\tilde{\mathbf{S}}_\ell$ expands $\mathbf{S}_\ell$ such that

$$\tilde{\mathbf{S}}_\ell = \begin{bmatrix}\mathbf{I}_C & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_\ell\end{bmatrix}.$$

Since deterministic elements are exogenous to the estimation of the hierarchy, the corresponding parameters are not regularized. Apart from a regression constant, deterministic elements can include statistical features such as trends or dummy variables, or simply predictors that, for one reason or another, are better represented outside of the feature hierarchy. All applications in this paper contain a deterministic element in the form of a regression constant.

The analogy of the HFR to a model average over level-specific regressions can furthermore be extended to obtain approximate standard errors of the parameter estimates. Since level-specific standard errors, $\hat{\text{se}}(\hat{\mathbf{w}}_\ell)$, are readily retrieved from the level-specific regressions, the average standard errors $\hat{\text{se}}(\hat{\boldsymbol{\beta}}_{\text{hfr}})$ can be obtained following Burnham and Anderson (2004), with

$$\hat{\text{se}}(\hat{\boldsymbol{\beta}}_{\text{hfr}}) = \sum_{\ell=1}^{L} \phi_\ell \sqrt{\hat{\text{se}}(\hat{\mathbf{w}}_\ell)^2 + (\hat{\mathbf{w}}_\ell - \hat{\hat{\mathbf{w}}}_{\boldsymbol{\phi}})^2}, \tag{3.18}$$

where the weighted average parameters $\hat{\hat{\mathbf{w}}}_{\boldsymbol{\phi}}$ are simply the HFR estimates $\hat{\boldsymbol{\beta}}_{\text{hfr}}$. It is important to note that for purposes of inference $\hat{\text{se}}(\hat{\boldsymbol{\beta}}_{\text{hfr}})$ are understated. For instance, the graph estimation error embedded in $\mathbf{S}$ is omitted entirely. Nonetheless, the standard errors provide valuable information about the average significance along the branch of each variable in the hierarchy, and can be useful to prune noise clusters and to inform sparse model selection, as illustrated in the following section. Once again, in the absence of shrinkage, with $\boldsymbol{\theta} = \mathbf{1}$, the standard errors $\hat{\text{se}}(\hat{\boldsymbol{\beta}}_{\text{hfr}})$ are equivalent to the standard errors of the OLS regression.

An additional tool in understanding the role of the optimal parameter graph and in empirically validating Assumption 1 is to examine the level-wise decomposition of the coefficient of determination. Letting the model fit up to the ℓth level be given by

$$\hat{y}_{\to\ell} = \sum_{i=0}^{\ell-1} \theta_{\ell-i} \mathbf{x} \hat{\mathbf{b}}_{\ell-i}, \tag{3.19}$$

the cumulative coefficient of determination can be defined in the usual manner, with

$$R^2_{\to\ell} = 1 - \frac{\sum_{i=1}^{N} ([\hat{y}_{\to\ell}]_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}. \tag{3.20}$$

When $\ell = L$, this simply results in the total R^2 of the HFR fit. However, the level-wise formulation in Eq. 3.20 also yields contributions, R^2_ℓ, of each individual level to the overall coefficient of determination, where $R^2_\ell = R^2_{\to\ell} - R^2_{\to\ell-1}$ and $\sum_\ell R^2_\ell = R^2$. A higher contribution by the initial levels suggests that few clusters explain a large proportion of the variance in y, while high contributions in lower levels are indicative of a higher degree of idiosyncratic variable information.

Fig. 3.6 illustrates this concept, by plotting R^2 contribution profiles of HFR estimates for two simulated data sets: (i) data containing clusters and (ii) data containing no clusters (and thus violating Assumption 1).[10]

[9] As an aside, the HFR can be made entirely sign-invariant, permitting negatively correlated predictors with a similar explanatory effect on y — albeit with opposite signs — to be clustered adjacently. This is achieved by using the absolute partial correlation matrix, $|\mathcal{D}_y|$. Such an approach is useful when the sign is not deemed to convey meaningful information, with only the absolute size of the coefficients being relevant.

[10] The data sets are reproduced from Simulations (a) and (a*) in Section 5.1.

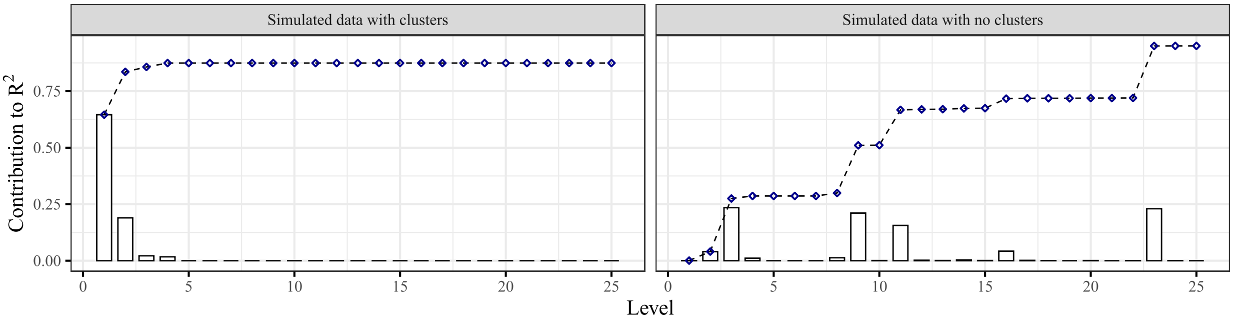


Fig. 3.6. Level-wise R^2 contribution profile of HFR estimation with clustered data (left panel) and unclustered data (right panel). The data is reproduced from Simulations (a) and (a*) in Section 5.1. Bars represent level-wise contributions to total R^2, the dashed line represents cumulative contributions. Only levels $\ell < 25$ are plotted.

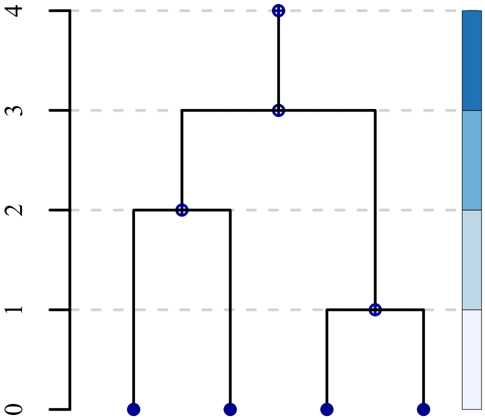


Fig. 3.7. Illustrative dendrogram of the level-wise decomposition of the HFR estimates with level-specific contributions on the right.

In the plots in Section 4, the level contributions are added to the dendrograms as bars with darker colors suggesting a larger contribution of that level, as illustrated in Fig. 3.7.

As a final issue, the discussion has thus far assumed $K < N - M$. When $K \geq N - M$, the level-specific regressions for all $\ell \geq N - M$ cannot be computed. Since the lowest levels group predictors with the highest similarity, the simplest remedy is to prune all levels where $\ell \geq N - M$. This leaves a total of $L = N - M - 1$ levels with no effect on the structure of the HFR, with the sole exception that the constraint in Eq. 3.10 substitutes $\kappa(N - M - 1)$ for κK:

$$\begin{aligned}\boldsymbol{\phi}_\kappa^* = \arg\min_{\boldsymbol{\phi}} & \left[\boldsymbol{\phi}^\top \mathbf{U}\boldsymbol{\phi} - \mathbf{V}^\top \boldsymbol{\phi}\right] \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\phi} \leq \mathbf{1} \quad \text{and} \\ & \boldsymbol{\phi}^\top \boldsymbol{\omega}\mathbf{1} = \kappa(N - M - 1).\end{aligned} \tag{3.21}$$

A complete implementation of the HFR algorithm is provided in the `hfr` package available on the Comprehensive R Archive Network (CRAN) for the statistical computing language R (Pfitzinger, 2023).

4. A case study: Determinants of economic growth

The HFR is useful both as a regression estimator and as a tool to garner insights into the effect structure underlying an estimated statistical model. In this section, I propose an analysis workflow that uses the HFR to understand an empirical problem and to obtain robust out-of-sample predictions. The data is taken from Sala-I-Martin et al. (2004), who in their seminal paper on the determinants of economic growth, compile a cross-country data set comprising GDP per capita growth rates between 1960-1996 for a sample of 88 countries, alongside 67 potential explanatory variables. The variables in-

Table 4.1
Distribution of prediction error (MSE) for GDP per capita growth based on 500 simulation runs. Prediction errors are multiplied by 1e4.

	HFR	Ridge	PLSR	PCR	Lasso	ElasticNet	AdaLasso	BACE	BEN
Min.	0.209	0.140	0.173	0.212	0.267	0.140	0.146	0.358	0.452
1st Qu.	0.942	0.963	1.008	1.032	1.196	1.072	1.421	1.122	1.114
Median	1.339	1.408	1.471	1.568	1.756	1.586	2.053	1.614	1.494
Mean	1.554	1.625	1.784	1.856	1.986	1.792	2.370	1.705	1.587
3rd Qu.	1.860	2.132	2.279	2.360	2.516	2.299	2.870	2.158	1.869
Max.	5.779	7.083	10.036	6.985	9.307	9.307	20.432	4.213	3.891

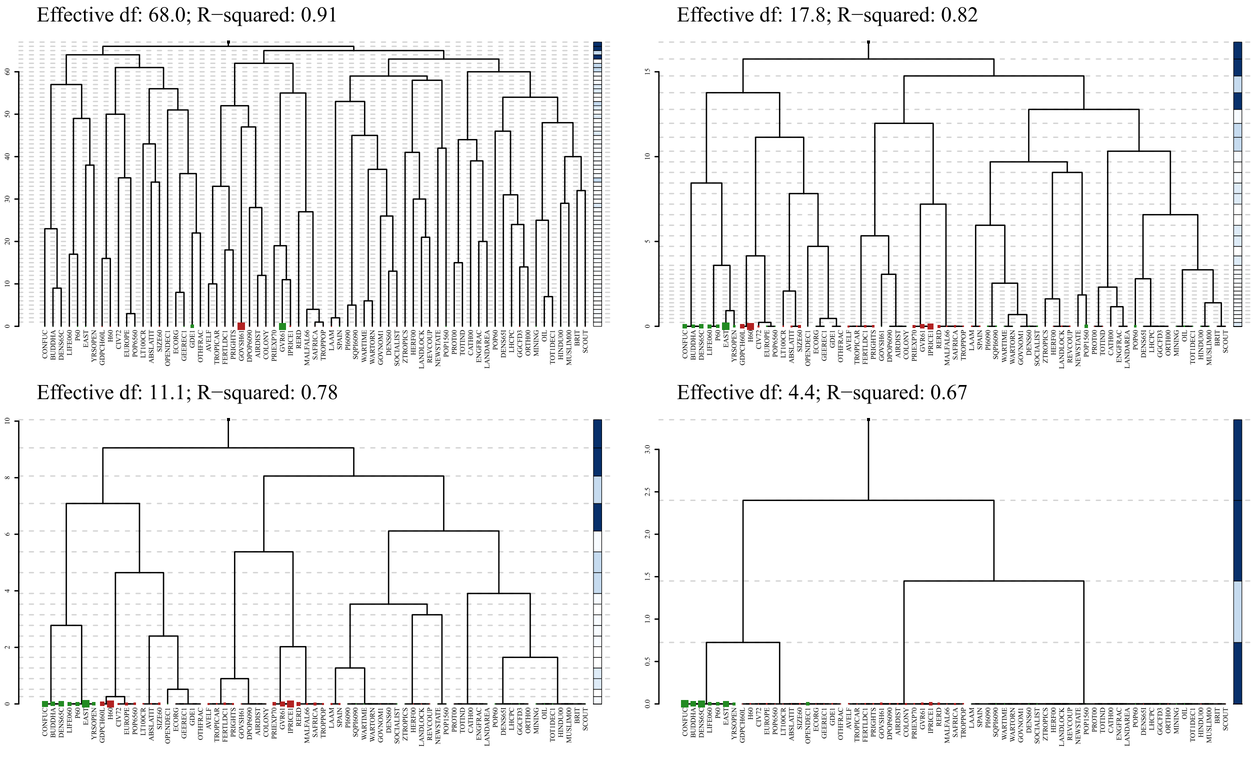


Fig. 4.1. Dendrograms for the HFR coefficients obtained from four different settings for κ. The top-left panel represents the unconstrained case ($\kappa = 1.00$), while the bottom-left panel is the hyperparameter that minimizes a 10-fold cross-validated MSE ($\kappa = 0.15$). The remaining panels are generated by $\kappa = 0.25$ (top) and $\kappa = 0.05$ (bottom).

clude measures that capture underlying socio-economic concepts such as institutional quality, demographics or geographical constraints. A description of the variables contained in the data set is provided in Table Appendix D.1.[11]

The data set has become a workhorse for testing high-dimensional regression techniques, particularly in the Bayesian literature (Eicher et al., 2011; Hofmarcher et al., 2011; Ley, 2008; Sala-I-Martin et al., 2004; Schneider and Wagner, 2012). The econometric techniques that have been employed include Bayesian model averaging, as well as various model selection and shrinkage methods such as the elastic net and lasso estimators. Since the data set comprises clusters of related measures of underlying concepts, it may be reasonable to expect a latent factor structure such as assumed by the HFR to offer a more efficient bias-variance trade-off than the sparsity assumption underpinning many of the penalized estimators applied in past research.

As a starting point, Fig. 4.1 depicts hierarchical graphs for the data set of economic growth determinants using 4 different settings of κ — the hyperparameter governing the size of the optimal graph. The unconstrained regression graph is plotted in the top-left panel, with a total height of 67 ($\kappa = 1$) and each level contributing to a maximum extent ($\boldsymbol{\theta} = \mathbf{1}$). The graph is highly complex, reflecting the dimensionality of the problem, and is difficult to interpret in a meaningful manner. The

[11] Since the HFR as well as benchmark methods require the ranges of the input variables to be similar, the 67 predictors in the data set are scaled to an interval of $[-1, 1]$. Dummy variables are normalized to a range of $[-0.5, 0.5]$. This is done to dampen the otherwise overstated effect of the dummy variables. The GDP per capita growth variable is not transformed to ensure that a comparison to previous research is possible.

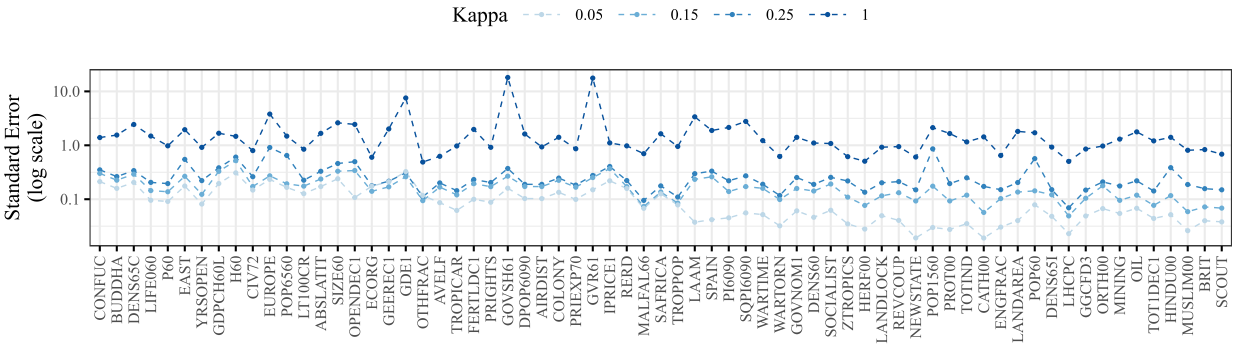


Fig. 4.2. Approximate standard errors of the HFR estimates using four different settings for κ (log scale). As the bias of the estimates increases with higher κ, the variance decreases. The standard errors represent weighted averages over the level-specific standard errors as described in Section 3.7.

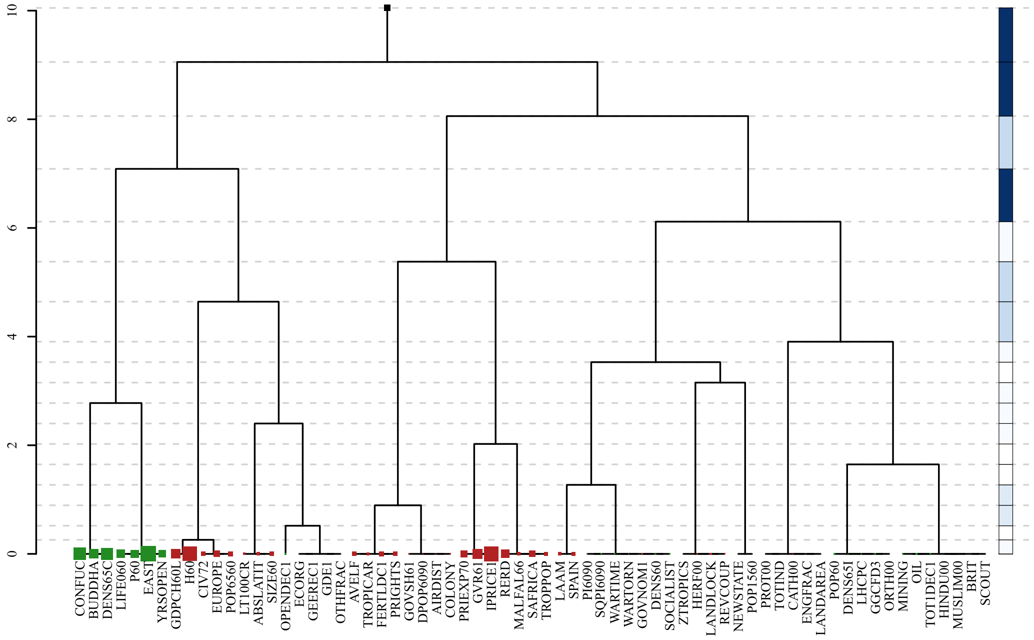


Fig. 4.3. Dendrogram for the HFR coefficients for the hyperparameter that minimizes a 10-fold cross-validated MSE. The size of the leaf nodes represents the magnitude of the parameter estimates and the color denotes the sign (green is positive, red is negative). The shaded bar on the right represents level-specific contributions to total R^2 with darker hues indicating a higher contribution.

regression coefficients themselves, which are represented by the leaf nodes, are estimated with substantial variance (see Fig. 4.2), highlighting the need for a regularized approach.

The remaining panels of Fig. 4.1 show different degrees of shrinkage, leading to successively simpler hierarchies. Each lower value of κ increases the strength of shrinkage (and hence the parameter bias), while in turn decreasing the variability of the estimates, as demonstrated in Fig. 4.2.

In contrast to the complex unconstrained structure, Fig. 4.3 displays the estimated optimal shrinkage tree for the regression. The height of the tree is 10.1, with $\nu_{\text{eff}} = 10.1 + 1$ determined using a 10-fold cross-validation procedure. The distance between the levels reflects the shrinkage weights $\boldsymbol{\theta}^*_\kappa$, and the vertical bar on the right is shaded based on the contribution of each level to the overall coefficient of determination of the HFR fit.

Fig. 4.3 suggests that the primary contribution to model fit is derived from the upper levels. Examining the level-wise contributions directly in Fig. 4.4 shows that only the first 18 levels contribute to the fitting process and the first four levels account for over 85% of the explained variation. The plot is analogous to scree plots produced for principal components regressions, with the summation over the level-specific contributions yielding the total R^2 of the HFR fit.

As illustrated in the bottom-right panel of Fig. 4.1, the first four hierarchical levels divide the sample into four latent signal factors that explain a significant portion of the response variation. The factors appear to identify regional or topical sub-clusters, as well as consolidated noise components. The first cluster contains variables that identify the East Asian re-

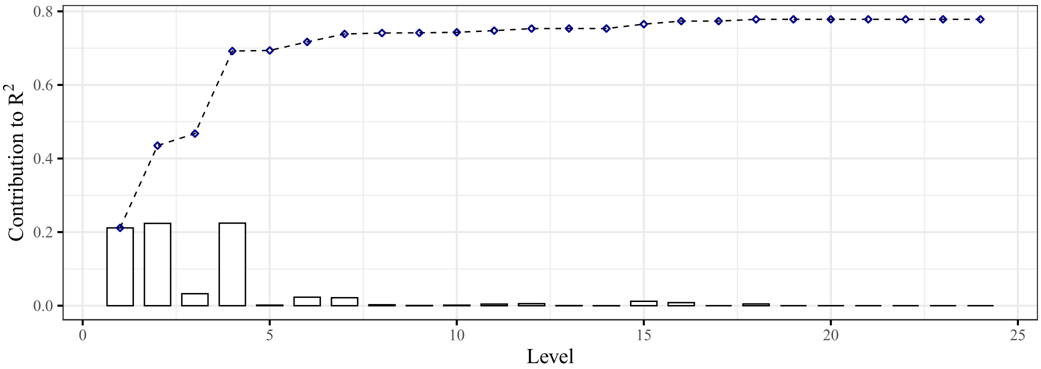


Fig. 4.4. Cumulative R^2 over the levels in the optimal hierarchy (dashed line), and level-specific contributions to R^2 (bars). Only levels $\ell < 25$ are plotted. The remaining levels do not contribute to the model fit.

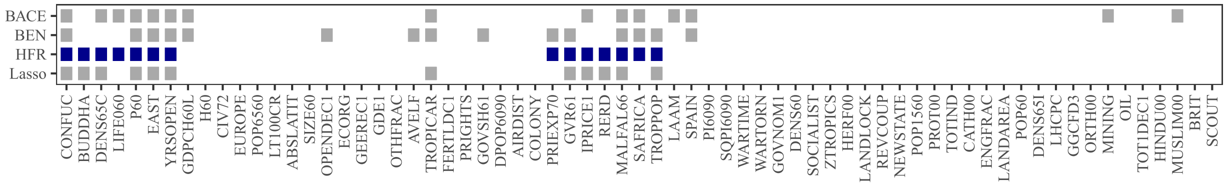


Fig. 4.5. Model selection using the approximate p-values of the HFR coefficients compared to BACE, BEN and lasso regressions.

gion (e.g. BUDDHA, CONFUC, EAST).[12] The second cluster appears to group mostly institutional quality measures and some related variables (e.g. H60, CIV72, OPENDEC1, ECORG). The third cluster groups variables that presumably identify developing economies (e.g. MALFAL66, SAFRICA, TROPPOP) and several closely related economic measures (e.g. RERD, IPRICE1, PRIEXP70). Finally, the fourth cluster contains a large group of variables with coefficients close to zero, suggesting that these measures represent primarily noise components.

The fact that the upper clusters enter with a much higher importance than their corresponding leaf nodes, may suggest that the common — as opposed to the idiosyncratic — information in the predictor groups determines growth disparities. For instance, rather than malaria prevalence entering as a growth determinant in its own right, the variable (MALFAL66) helps to identify an underlying geographic factor that drives economic growth.

Examining the individual growth drivers more closely, Fig. 4.5 displays the most important variables identified by Sala-I-Martin et al. (2004) (BACE) and Hofmarcher et al. (2011) (BEN), as well as all HFR coefficients with an indicative p-value $<$ 0.05.[13] As an auxiliary comparison, the model selected using a lasso estimator is also displayed.[14] The HFR identifies a total of 14 growth determinants grouped into two blocks: those associated with cluster one and those associated with cluster three. The variable set closely resembles related studies (with 12 of 14 overlapping drivers), but reflects the clustering inherent to the HFR. The model selected using the lasso is almost identical to the HFR, with all but one of the growth determinants taken from the two relevant effect clusters discovered by the HFR.

A key consideration for the validity of the uncovered model and the quality of the HFR estimates is the method's predictive performance. In order to assess this systematically, I employ a sampling setup closely resembling Hofmarcher et al. (2011). Observations are randomly sampled to form training, validation and testing sets containing 68/10/10 observations, respectively.[15] Parameters are estimated using the training sample, hyperparameters are determined via a grid search minimization of the validation mean squared error (MSE) and the performance is calculated as the test sample MSE. Samples are drawn in 500 iterations with hyperparameters determined independently in each run.

Fig. 4.6 plots the MSE and the average rank for the HFR and a panel of benchmark methods. The benchmark methods include penalized regressions in the form of the ridge regression, lasso, adaptive lasso (AdaLasso) and elastic net, latent variable regressions in the form of PCR and PLSR, and finally OLS.[16] In addition, Table 4.1 displays the distribution of the

[12] See Table Appendix D.1 for the full variable descriptions.

[13] The p-value is calculated using average standard errors as described in Section 3.7 with the residual degrees of freedom given by $N - \nu_{\text{eff}}$.

[14] As for the HFR, the lasso penalty is determined using a 10-fold cross-validation approach.

[15] These proportions are roughly equivalent to those used in Hofmarcher et al. (2011), but with the addition of a validation sample, which is obtained by reducing the size of both the training and testing samples slightly.

[16] Ridge, lasso, AdaLasso and elastic net are implemented using the `glmnet`-package in the statistical computing language R, described in Friedman et al. (2010). For a discussion of the AdaLasso, see Zou (2006). PCR and PLSR are implemented using the `pls`-package in the statistical computing language R, described in Mevik and Wehrens (2019). Hyperparameters include κ for the HFR, the size of the penalty (λ) for the penalized estimators (ridge, lasso, AdaLasso, elastic net), the mixing parameter (α) for the elastic net, and the number of latent components for the PCR and PLSR.

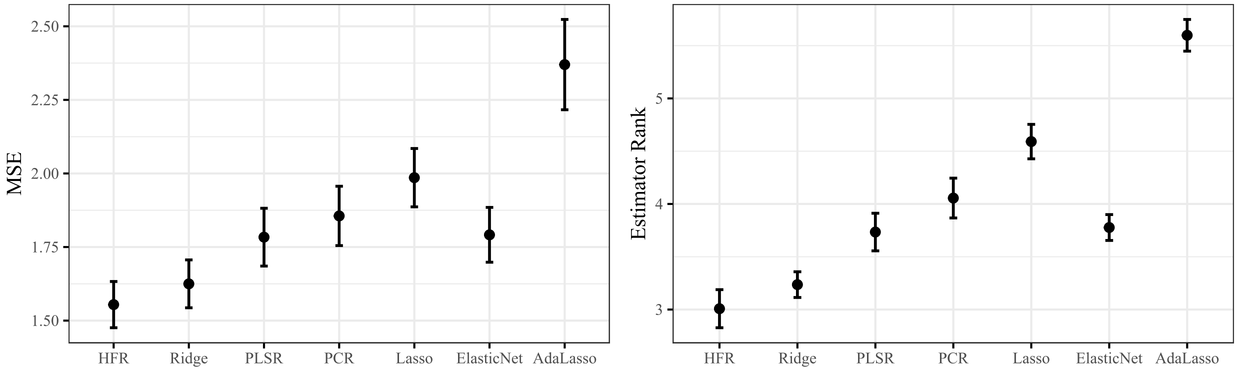


Fig. 4.6. Comparison of prediction accuracy of HFR, Ridge, PLSR, PCR, Elastic Net, Lasso, and AdaLasso for GDP per capita growth from 1960-1996. MSE (left panel), rank of estimators (right panel). Statistics plotted as mean and 95% confidence interval based on 500 random training, validation and testing samples. Prediction errors are multiplied by 1e4.

MSEs alongside the results of the BACE and the BEN. The estimation of BACE and BEN is not replicated, but the results are taken directly from the table presented in Hofmarcher et al. (2011), page 10.

The HFR outperforms all benchmark methods, with the ridge regression achieving the highest mean accuracy among the panel of non-Bayesian benchmarks in Fig. 4.6, followed by the PLSR. When compared to the performance of the BACE and BEN models, the HFR is again found to achieve lower mean and median prediction errors. The results provide justification for the approach taken by the HFR, suggesting that the aggregation of growth determinants into a low-dimensional set of latent factors is indeed appropriate.

In sum, the HFR offers a dual benefit: (i) it generates robust out-of-sample predictions, while (ii) the parsimonious hierarchy, in which the estimates are embedded, produces meta-insights about the underlying latent signals that explain observed response variation. In the case of the determinants of economic growth, several regional and topical sub-clusters may suffice to offer robust explanations of observed growth disparities. The following section tests the generality of the observed predictive accuracy under simulated conditions.

5. Simulations

I use two sets of simulations, largely replicated from related work, to compare the performance of the HFR to similar methods. In Section 5.1, I examine a regression problem with a mixture of sparsity and grouping, and compare the HFR to other group shrinkage methods. In Section 5.2, I explore more generally the performance of the HFR benchmarked against a panel of the most commonly used regularized regression techniques across a spectrum of different regression problems.

Data is simulated from the true model

$$y = \mathbf{x}\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Observations are divided into training, validation and testing samples, where the training sample is used to estimate the models, the validation MSE determines optimal hyperparameters using a grid search, and the testing sample is used for performance evaluation. Model performance is assessed by calculating the MSE over the testing sample. Sample sizes are denoted by $\cdot/\cdot/\cdot$, where the dots represent training, validation and testing samples, respectively. In each case, the results of 500 simulation runs are plotted.

5.1. Group shrinkage

Simulation (a) is based on Zou and Hastie (2005), who study the effect of grouped predictors. The simulation contains a mixture of grouped predictors and noise predictors and is therefore a grouped feature selection task. The sample consists of 50/50/400 observations and 40 predictors with

$$\boldsymbol{\beta} = (\underbrace{3, ..., 3}_{5}, \underbrace{4, ..., 4}_{5}, \underbrace{5, ..., 5}_{5}, \underbrace{0, ..., 0}_{25}),$$

$\sigma^2 = 15$, and $\mathbf{x}$ generated as follows (with $\epsilon_i^x \sim \mathcal{N}(0, 0.01)$):

$$\begin{aligned}
x_i &= \xi_1 + \epsilon_i^x, \quad \xi_1 \sim \mathcal{N}(0,1), \quad i = 1, ..., 5,\\
x_i &= \xi_2 + \epsilon_i^x, \quad \xi_2 \sim \mathcal{N}(0,1), \quad i = 6, ..., 10,\\
x_i &= \xi_3 + \epsilon_i^x, \quad \xi_3 \sim \mathcal{N}(0,1), \quad i = 11, ..., 15,\\
x_i &\sim \mathcal{N}(0,1), \quad i = 16, ..., 40.
\end{aligned}$$

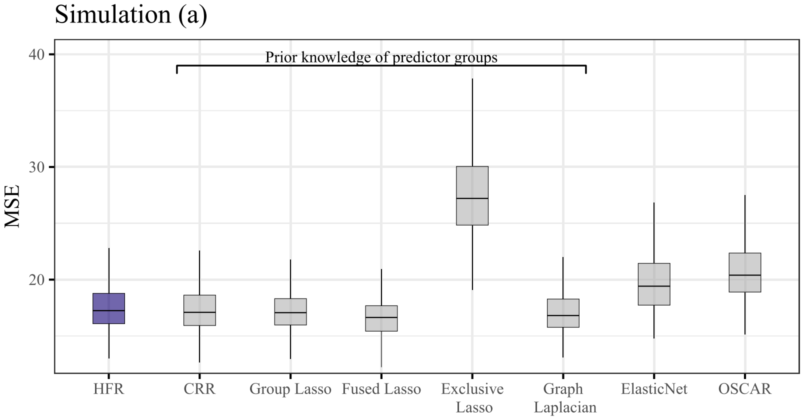


Fig. 5.1. Comparison of prediction accuracy of hierarchical feature regression (HFR), CRR, Group Lasso, Fused Lasso, Exclusive Lasso, Graph Laplacian, Elastic Net and OSCAR for simulation (a).

The simulation is well-suited to exploring the performance of group shrinkage estimators. Particularly interesting, in this context, is the comparison between structured sparsity, group target shrinkage and graph-based estimators, the latter of which is a hybrid between the former two classes of methods. I include the group lasso (structured sparsity), the clustered ℓ_2-norm from Eq. 2.2 (denoted clustered ridge regression, CRR), as well as the fused lasso and graph Laplacian (graph-based methods). All four of these methods require prior knowledge of predictor groups.[17]

In addition, I calculate the performance of group shrinkage methods that do not require prior knowledge, including — apart from the HFR — the OSCAR and the elastic net. Note that the elastic net is introduced in Zou and Hastie (2005) as a method to deal with grouped data. Finally, the exclusive lasso is added to illustrate the effect of an incorrect assumption about the DGP (the exclusive lasso allows sparsity within each group).[18]

Fig. 5.1 shows virtually identical performance for HFR and the group shrinkage estimators, despite the fact that the latter methods incorporate prior knowledge of the structure of the DGP. An incorrect assumption about the DGP, in the case of the exclusive lasso, leads to comparatively poor performance.

Fig. 5.2 studies the issue of misspecification. The upper panel displays variants of CRR, group and fused lasso, and graph Laplacian, where group compositions are determined using hierarchical clustering (instead of being given) as proposed, for instance, in Bühlmann et al. (2013), and the number of groups (G) is incorrectly specified. Recall that the CRR is conceptually equivalent to the HFR with only a single level. By including all levels, the HFR does not require prior specification of G.

The lower panel of Fig. 5.2 adjusts Simulation (a), removing the grouping structure entirely from the data with $x_i \sim \mathcal{N}(0, 1),\ i = 1, \ldots, 40$ (violating Assumption 1). The CRR and fused lasso are computed once again using hierarchical clustering (which is spurious in the uncorrelated case) and $G = 4$, while the group lasso and graph Laplacian, by way of illustration, are calculated with prior knowledge of true group compositions and the true Laplacian matrix, respectively.

The plots convey two observations: (i) All four group shrinkage methods perform considerably more poorly when the number of clusters is misspecified (i.e. the cases of $G = 2$ and $G = 3$ in the upper panel of Fig. 5.2); (ii) When Assumption 1 is violated, the HFR and other methods that rely on the correlation structure, such as the CRR with hierarchical clustering, perform poorly. Estimators such as the elastic net and OSCAR are somewhat less sensitive to the correlation structure. Finally, the availability of exact prior grouping information (group lasso and graph Laplacian in the lower panel of Fig. 5.2) leads to best performance.

5.2. Comparison to common regularized regressions

Since the graph structure of the HFR is intrinsic to the data, this section compares it to other commonly used regularization methods, several of which encourage intrinsic grouping.

Simulation (b) is taken from Tibshirani (1996), where it was originally used to demonstrate the performance of the ridge regression. True parameter values are set to $\beta_j = 0.85,\ \forall\, j = 1, \ldots, K$, with $K = 8$. The sample size is 20/20/200, $\sigma^2 = 3$ and the pairwise correlation between x_i and x_j is $0.5^{|i-j|}$.

[17] The methods are implemented in the statistical computing language R using Yang et al. (2020) (group lasso), Allaire and Tang (2022) (CRR), Taylor and Tibshirani (2022) (fused lasso) and Chen and Chen (2015) (graph Laplacian). Both group lasso and CRR have a single hyperparameter, λ, which governs the size of the penalty as in the case of the lasso regression. The fused lasso and graph Laplacian have an additional hyperparameter governing the trade-off between sparsity and smoothness penalty terms.

[18] The exclusive lasso (Zhou et al., 2010) is implemented using the `ExclusiveLasso` package in R (Weylandt and Campbell, 2018) with hyperparameter λ. The OSCAR (Bondell and Reich, 2008) is implemented using the `lqa` package in R (Ulbricht, 2012). The method has two hyperparameters, λ and c. The elastic net is implemented using the `glmnet` package in R Friedman et al. (2010) with hyperparameters λ, as well as the mixing parameter α.

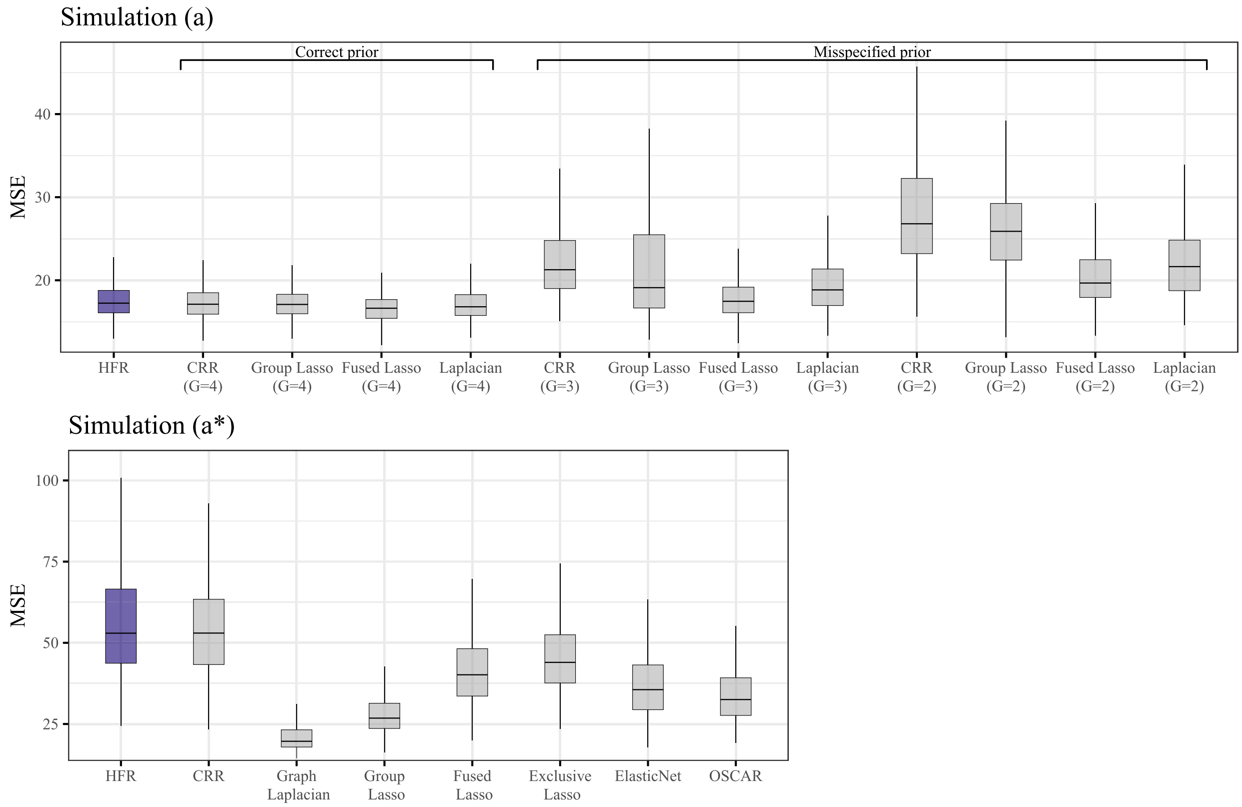


Fig. 5.2. Comparison of prediction accuracy of hierarchical feature regression (HFR), Group Lasso, Fused Lasso, Graph Laplacian and CRR for different grouping assumptions (G) (upper panel). Comparison of prediction accuracy of HFR, CRR, Group Lasso, Fused Lasso, Graph Laplacian, Exclusive Lasso, Elastic Net and OSCAR for data with independent predictors (lower panel).

Simulation (c) is again based on Tibshirani (1996) and is a sparse regression used to illustrate the lasso's ability of eliminating noise features. There are 40 predictors with parameters set to

$$\boldsymbol{\beta} = (\underbrace{0, ..., 0}_{10}, \underbrace{2, ..., 2}_{10}, \underbrace{0, ..., 0}_{10}, \underbrace{2, ..., 2}_{10}).$$

As before the pairwise correlation between x_i and x_j is $0.5^{|i-j|}$, and $\sigma^2 = 8$. The sample size is set to 100/100/400. Since the DGP is sparse, the task is likely to be solved well with a lasso, adaptive lasso or elastic net. While the setup is changed slightly from Tibshirani (1996), the original simulation is replicated in a more detailed exploration in Section 5.3.

Simulation (d) is designed to test predictive performance in the presence of latent factors. The sample consists of 20/20/200 observations. Simulation (d) draws from a true model $y = \boldsymbol{f}\boldsymbol{\beta} + \epsilon$ where $\boldsymbol{f} = \begin{bmatrix} f_1 & \cdots & f_4 \end{bmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} 1.0, 1.5, 2.0, 1.5 \end{pmatrix}$, $\sigma^2 = 3$ and the pairwise correlation between f_i and f_j is $0.5^{|i-j|}$. Unlike the previous cases, I assume $\mathbf{x}$ contains noisy measures of the unobserved latent factors $\boldsymbol{f}$, such that (with $\epsilon_i^x \sim \mathcal{N}(0, 1)$):

$$\begin{aligned} x_i &= f_1 + f_2 + \epsilon_i^x, \quad i = 1, 2, \\ x_i &= f_2 + f_3 + \epsilon_i^x, \quad i = 3, 4, \\ x_i &= f_3 + f_4 + \epsilon_i^x, \quad i = 5, 6, \\ x_i &= f_4 + f_1 + \epsilon_i^x, \quad i = 7, 8. \end{aligned}$$

The PCR and PLSR are expected to outperform other regularized regressions in this example.

Fig. 5.3 plots the model accuracy for Simulations (a) to (d), comparing the HFR to the same panel of commonly used regularization methods described in Section 4. The HFR outperforms or closely matches the benchmarks in all simulations. Good performance in the cases when no predictor groups exist in the true DGP (Simulations (b) & (c)), or when an overlapping grouping structure exists (Simulation (d)) illustrate the versatility of the HFR in estimating robust parameters. The feature selection tasks (Simulations (a) & (c)) demonstrate how the ability to group noise features can lead to good performance even when compared to methods that explicitly perform variable selection, such as the lasso, adaptive lasso and elastic net regressions.

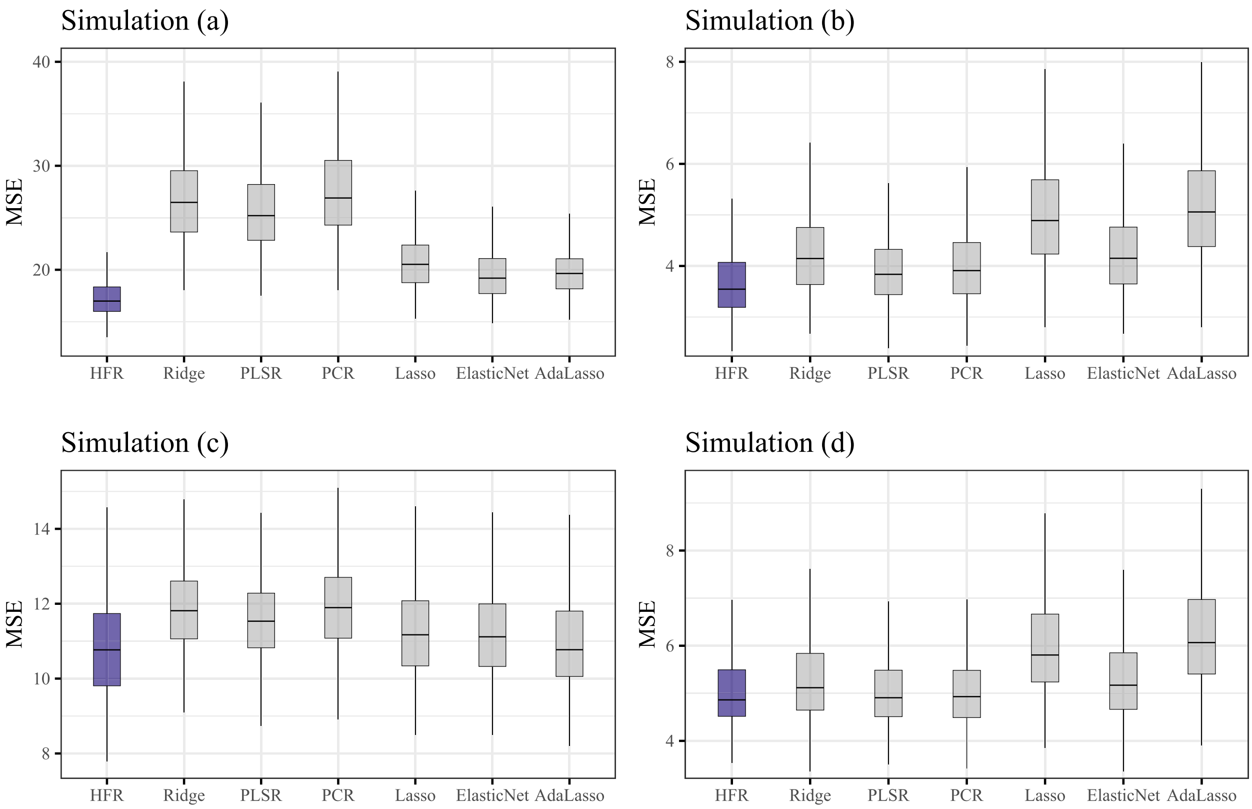


Fig. 5.3. Comparison of prediction accuracy of hierarchical feature regression (HFR), Ridge, PLSR, PCR, Elastic Net, Lasso and AdaLasso for simulations (a)-(d).

Table 5.1
Prediction accuracy (median MSE) for simulations (a)-(d) based on 500 simulation runs. Standard errors in parantheses. Standard errors are calculated using 500 bootstrap resamplings of the estimated MSE. In each case the two best methods are highlighted.

	Sim. (a)	Sim. (b)	Sim. (c)	Sim. (d)
HFR	**16.988** (0.084)	**3.546** (0.023)	**10.767** (0.068)	**4.862** (0.039)
Ridge	26.478 (0.265)	4.146 (0.06)	11.813 (0.056)	5.118 (0.05)
PLSR	25.204 (0.182)	**3.837** (0.041)	11.533 (0.062)	**4.907** (0.055)
PCR	26.902 (0.292)	3.909 (0.058)	11.896 (0.093)	4.93 (0.029)
Lasso	20.521 (0.134)	4.89 (0.054)	11.17 (0.053)	5.805 (0.056)
ElasticNet	**19.193** (0.154)	4.151 (0.055)	11.116 (0.045)	5.171 (0.048)
AdaLasso	19.641 (0.118)	5.059 (0.055)	**10.772** (0.088)	6.067 (0.074)
OLS	83.194 (2.157)	5.4 (0.111)	13.413 (0.078)	7.318 (0.133)

Table 5.1 summarizes the results of the simulations including bootstrap standard errors for the median MSE performance metrics.

5.3. *Partial correlations and sparsity*

The above results suggest good comparative accuracy in high-dimensional feature selection tasks, which can be traced to the HFR's ability to group noise features efficiently. Nonetheless, a more complex sparse DGP can make meaningful graph estimation based on the bivariate partial correlation profile impossible, as illustrated in the following variation on Simulation (c):

Simulation (c^*) is taken from Tibshirani (1996) and is identical to Simulation (c) with 40 predictors and a sparse DGP. Unlike in the previous case, all predictors have a pairwise correlation of 0.5, and $\sigma^2 = 15$. The high pairwise correlation between noise and non-noise predictors ensures that bivariate partial correlations are insufficient to separate signal from noise.

Fig. 5.4 plots the simulation results for Simulations (c) and (c^*) with the HFR calculated using bivarate partial correlations ($\mathcal{D}_y^{(b)}$), as well as Ledoit and Wolf (2003) shrinkage estimates of the partial correlation ($\mathcal{D}_y^{(s)}$). While the shrinkage estimates perform well in both cases, the estimates based on bivariate partial correlations produce a result similar to the PLSR and substantially less accurate than the feature selection methods in Simulation (c^*).

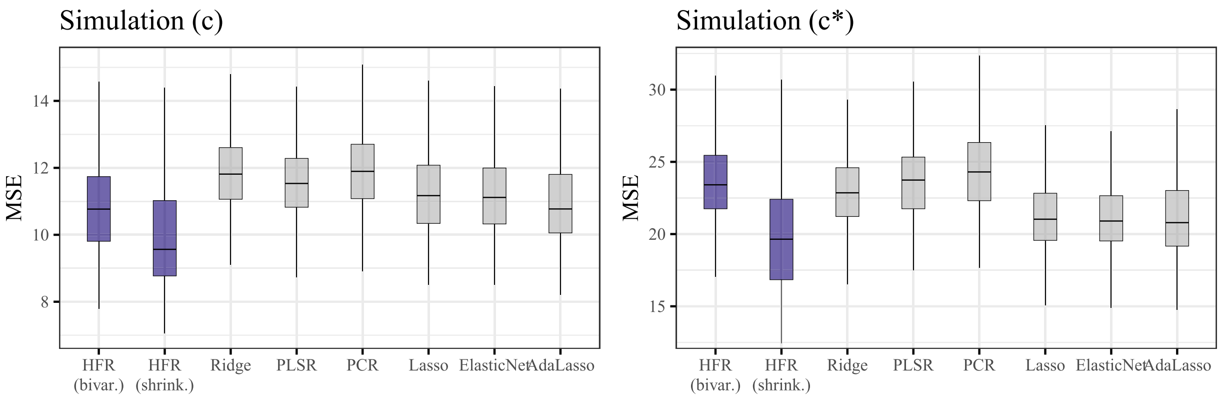


Fig. 5.4. Comparison of prediction accuracy of hierarchical feature regression (HFR), Ridge, PLSR, PCR, ElasticNet, Lasso and AdaLasso for simulations (c) and (c*).

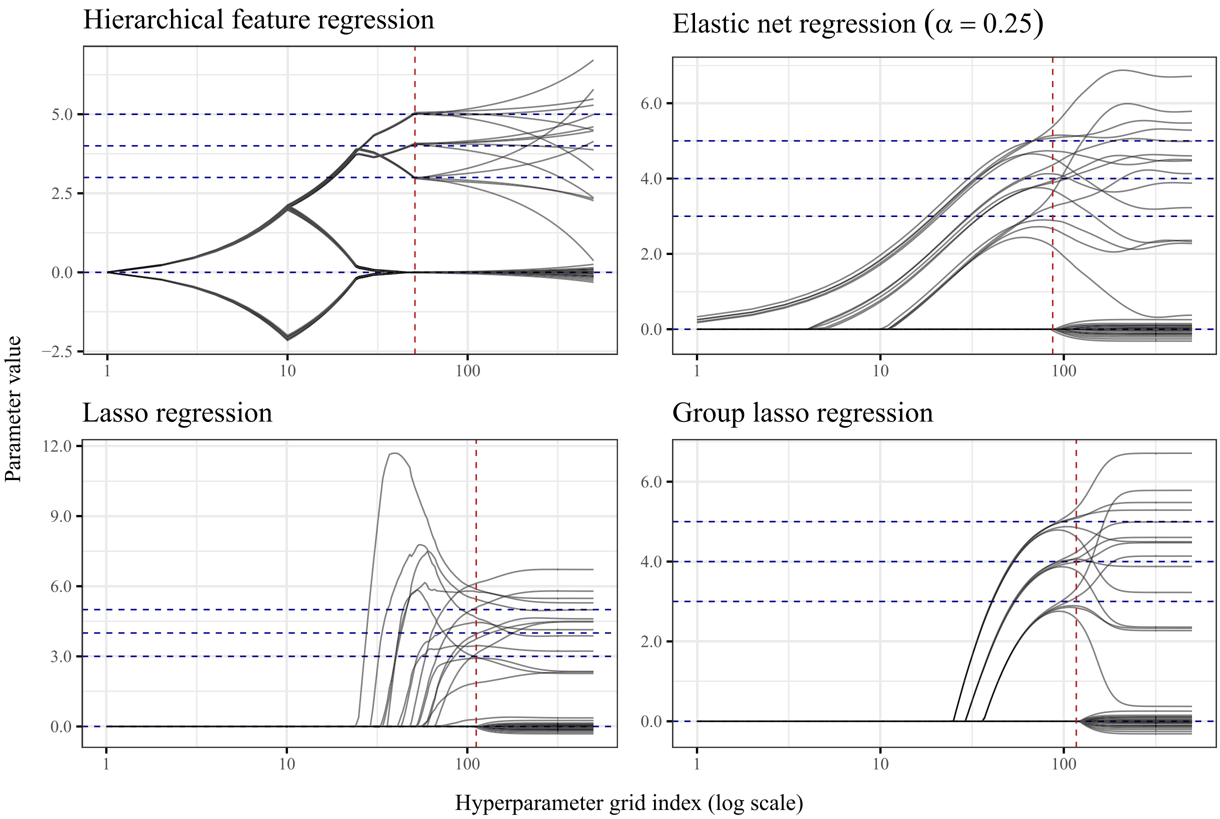


Fig. 5.5. Trace plots for HFR, elastic net, lasso and group lasso for $K = 40$ predictors. The hyperparameter values are κ for the HFR, and λ for the remaining methods. True parameter values are represented by horizontal dashed lines. Optimal hyperparameter values (determined using a 10-fold cross validation) are indicated by the vertical dashed line.

Despite this result, the drawbacks of using high-dimensional shrinkage estimates of the partial correlations (e.g. additional estimation risk, computational overhead) outweigh the benefits unless strong reason exists to suspect a complex effect structure. In such a case, however, an explicit feature selection method like the lasso regression may well be the preferred methodology. While the HFR can obtain robust estimates on noise features by grouping them, it does not explicitly select a sparse model and is therefore not a natural starting point for the analysis of a presumably complex sparse DGP.

5.4. Trace plots

Finally, I examine more closely the shrinkage behavior of the HFR by drawing trace plots for $\hat{\boldsymbol{\beta}}_{\text{hfr}}$ in Fig. 5.5 (top-left panel) using the setup in Simulation (a), with $K = 40$ predictors. The plot illustrates how parameter estimates are drawn towards group targets as κ decreases. The estimates are eventually shrunken towards zero for very small values of κ. When κ is sufficiently small, noise and non-noise predictors are captured in a single cluster, with the estimates on noise predictors diverging from zero. This reflects the lack of a sparsity objective.

By way of comparison, Fig. 5.5 (bottom and upper right panels) draws trace plots for the elastic net, lasso and group lasso estimators using the same regression problem as above. The HFR achieves grouping of estimates and reduction of noise with similar efficiency as the group lasso, however, without inducing sparsity directly. The elastic net, unlike the lasso, also induces grouping, but somewhat less efficiently than the HFR or group lasso, and generally introduces some attenuation bias to the estimates (true parameter values are indicated by dashed lines in Fig. 5.5).

6. Concluding remarks

Prediction tasks with high-dimensional multicollinear predictor sets are challenging for least squares based fitting procedures, and a large, productive literature exists advancing various regularized approaches to addressing the issue. The HFR is a novel contribution to this body of knowledge, presenting a method of shrinking coefficients towards group targets along the edges of an optimal predictor graph. Given a hyperparameter, which is conveniently interpreted as the effective model size and bounded between 0 and 1, the HFR is able to estimate both a supervised graph, as well as the optimal regularized coefficients associated with that graph.

The characteristics of the HFR make it particularly well-suited to regression applications with data that reflect an underlying hierarchical or grouped structure, with multiple predictors measuring similar latent concepts. Examples include high-dimensional modeling in econometric analysis (e.g. nowcasting of macroeconomic indicators) or in finance (e.g. multi-factor asset pricing), applications similar to the gene selection problem discussed in Zou and Hastie (2005), and numerous high-dimensional problems across other statistical domains. The ability to plot the estimated hierarchy and explore the effect of individual clusters or levels in the regression provides a wealth of auxiliary insights into the underlying effect structure.

Both the empirical case study and the simulations presented in this paper suggest that the HFR provides an interesting complement to widely used regularized regression algorithms such as the lasso or PLS regressions, as well as to common group shrinkage methods. The HFR achieves lower out-of-sample prediction errors than a panel of benchmark methods across a spectrum of different regression tasks, making it interesting both in terms of its performance as well as its versatility. The method can be thought of as a hybrid between a penalized regression with a clustered penalty similar to the clustered ℓ_2-norm and a supervised latent factor regression, with some benefits of both classes of algorithms, and potentially good performance across a wider range of data generating processes.

Declaration of competing interest

I hereby confirm that no competing interests are present in the research and development of the enclosed manuscript.

Appendix A. Proof of Proposition 1

Proposition 1 can be shown to hold by demonstrating the equivalency to ordinary least squares coefficients. The proposition defines

$$\hat{\boldsymbol{\beta}} = \mathbf{S}^\top \mathbf{Q}_{zz}^{-1} \mathbf{Q}_{zy}, \tag{A.1}$$

where

$$\mathbf{Q}_{zz} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Q}_{L1} & \mathbf{Q}_{L2} & \cdots & \mathbf{Q}_{LL} \end{bmatrix} \quad \text{and} \quad \mathbf{Q}_{zy} = \begin{bmatrix} \mathbf{Q}_{1y} \\ \mathbf{Q}_{2y} \\ \vdots \\ \mathbf{Q}_{Ly} \end{bmatrix},$$

with $\mathbf{Q}_{ij} = \mathbf{z}_i^\top \mathbf{z}_j$ and $\mathbf{Q}_{iy} = \mathbf{z}_i^\top y$, for any $i, j \in [1, L]$.

Eq. A.1 can be restated by splitting the lower triangular matrix into its diagonal ($\boldsymbol{\Lambda}$) and strictly triangular ($\mathbf{Q}_{zz}^*$) components,

$$\hat{\boldsymbol{\beta}} = \mathbf{S}^\top \left(\boldsymbol{\Lambda} + \mathbf{Q}_{zz}^*\right)^{-1} \mathbf{Q}_{zy}. \tag{A.2}$$

Expanding Eq. A.2 further yields (where $\mathbf{I}$ represents a diagonal matrix)

$$\hat{\boldsymbol{\beta}} = \mathbf{S}^\top \left[\boldsymbol{\Lambda}\left(\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{Q}_{zz}^*\right)\right]^{-1} \mathbf{Q}_{zy} \tag{A.3}$$

and

$$= \mathbf{S}^\top \left(\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{Q}_{zz}^*\right)^{-1} \boldsymbol{\Lambda}^{-1} \mathbf{Q}_{zy}. \tag{A.4}$$

Now use the well-known algebraic identity, $(1+x)\left(\sum_{j=0}^{m}(-x)^j\right) = 1-(-x)^{m+1}$, letting $x = \mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz}$ and $m = L-1$, to obtain

$$(\mathbf{I}+\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz})\left(\sum_{j=0}^{L-1}(-\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz})^j\right) = \mathbf{I}-(-\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz})^L. \tag{A.5}$$

Given the properties of strictly triangular matrices, $\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz}$ must be strictly triangular and nilpotent, with $\left(\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz}\right)^L = \mathbf{0}$. It follows, therefore, that

$$(\mathbf{I}+\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz})\left(\sum_{j=0}^{L-1}(-\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz})^j\right) = \mathbf{I} \tag{A.6}$$

and

$$(\mathbf{I}+\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz})^{-1} = \left(\sum_{j=0}^{L-1}(-\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz})^j\right). \tag{A.7}$$

Substituting Eq. A.7 back into Eq. A.4 yields

$$\hat{\boldsymbol{\beta}} = \mathbf{S}^\top\left(\sum_{j=0}^{L-1}(-\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz})^j\right)\mathbf{\Lambda}^{-1}\mathbf{Q}_{zy}. \tag{A.8}$$

In order to finalize this proof, it is necessary to examine the matrices $\mathbf{S}^\top\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz}$ and $\mathbf{S}^\top\mathbf{\Lambda}^{-1}\mathbf{Q}_{zy}$ from Eq. A.8, beginning with the former:

$$\mathbf{S}^\top\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz} = \begin{bmatrix}\mathbf{S}_1\\ \mathbf{S}_2\\ \vdots\\ \mathbf{S}_L\end{bmatrix}^\top \begin{bmatrix}\mathbf{Q}_{11} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}_{22} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_{LL}\end{bmatrix}^{-1} \begin{bmatrix}\mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{Q}_{21} & \mathbf{0} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{Q}_{L1} & \mathbf{Q}_{L2} & \cdots & \mathbf{0}\end{bmatrix}. \tag{A.9}$$

Now, let $\mathbf{S}_{\mathbf{c}^{(j)}}$ be the matrix $\mathbf{S}$ multiplied by scalar values, such that

$$\mathbf{S}_{\mathbf{c}^{(j)}} = \begin{bmatrix}c_1^{(j)}\mathbf{S}_1\\ c_2^{(j)}\mathbf{S}_2\\ \vdots\\ c_L^{(j)}\mathbf{S}_L\end{bmatrix}.$$

Furthermore, let $\mathbf{c}^{(j)} = \{c_\ell^{(j)}\}_{\ell=1,\dots,L}$ and let $\mathbf{c}^{(0)} = (1, \dots, 1)$ be a vector of ones, so that $\mathbf{S} = \mathbf{S}_{\mathbf{c}^{(0)}}$.

It is now possible to write

$$\mathbf{S}^\top_{\mathbf{c}^{(j)}}\mathbf{\Lambda}^{-1}\mathbf{Q}^*_{zz} = \begin{bmatrix}c_1^{(j)}\mathbf{S}_1\\ c_2^{(j)}\mathbf{S}_2\\ \vdots\\ c_L^{(j)}\mathbf{S}_L\end{bmatrix}^\top \begin{bmatrix}\mathbf{Q}_{11} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}_{22} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_{LL}\end{bmatrix}^{-1} \begin{bmatrix}\mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{Q}_{21} & \mathbf{0} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{Q}_{L1} & \mathbf{Q}_{L2} & \cdots & \mathbf{0}\end{bmatrix} \tag{A.10}$$

and

$$= \begin{bmatrix}c_1^{(j)}\mathbf{S}_1\\ c_2^{(j)}\mathbf{S}_2\\ \vdots\\ c_L^{(j)}\mathbf{S}_L\end{bmatrix}^\top \begin{bmatrix}\mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{Q}_{22}^{-1}\mathbf{Q}_{21} & \mathbf{0} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{Q}_{LL}^{-1}\mathbf{Q}_{L1} & \mathbf{Q}_{LL}^{-1}\mathbf{Q}_{L2} & \cdots & \mathbf{0}\end{bmatrix}. \tag{A.11}$$

Using the simple trick that $\mathbf{S}_i^\top(\mathbf{z}_i^\top\mathbf{z}_i)^{-1}\mathbf{z}_i^\top\mathbf{z}_j = \mathbf{S}_j^\top \ \forall\, i > j$, Eq. A.11 collapses to

$$\mathbf{S}^\top_{\mathbf{c}^{(j)}}\boldsymbol{\Lambda}^{-1}\mathbf{Q}^*_{zz} = \begin{bmatrix} \sum_{i=2}^{L} c_i^{(j)}\mathbf{S}_1 \\ \sum_{i=3}^{L} c_i^{(j)}\mathbf{S}_2 \\ \vdots \\ c_L^{(j)}\mathbf{S}_{L-1} \\ 0\mathbf{S}_L \end{bmatrix}^\top = \mathbf{S}^\top_{\mathbf{c}^{(j+1)}}, \tag{A.12}$$

where $\mathbf{c}^{(j+1)} = \{c_\ell^{(j+1)}\}_{\ell=1,\dots,L}$ and

$$c_\ell^{(j+1)} = \begin{cases} \sum_{i=\ell+1}^{L} c_i^{(j)} & \text{when } \ell < L \\ 0 & \text{otherwise.} \end{cases}$$

It follows that, with $\mathbf{c}^{(0)}$ given as a vector of ones, $\mathbf{c}^{(j)}$ represents the recursive definition of a binomial coefficient, with $c_\ell^{(j)} = \binom{L-\ell}{j}$.

Note furthermore that Eq. A.12 must generalize iteratively to any value of j, so that

$$\mathbf{S}^\top_{\mathbf{c}^{(0)}}(\boldsymbol{\Lambda}^{-1}\mathbf{Q}^*_{zz})^j = \mathbf{S}^\top_{\mathbf{c}^{(j)}}. \tag{A.13}$$

Thus, Eq. A.8 becomes

$$\hat{\boldsymbol{\beta}} = \left(\sum_{j=0}^{L-1}(-1)^j\mathbf{S}^\top_{\mathbf{c}^{(j)}}\right)\boldsymbol{\Lambda}^{-1}\mathbf{Q}_{zy}. \tag{A.14}$$

Examining now the second matrix, $\mathbf{S}^\top\boldsymbol{\Lambda}^{-1}\mathbf{Q}_{zy}$, this collapses simply to a sum of level-specific unconditional weights:

$$\mathbf{S}^\top\boldsymbol{\Lambda}^{-1}\mathbf{Q}_{zy} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_L \end{bmatrix}^\top \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_{LL} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Q}_{1y} \\ \mathbf{Q}_{2y} \\ \vdots \\ \mathbf{Q}_{Ly} \end{bmatrix}, \tag{A.15}$$

$$= \sum_{\ell=1}^{L}\mathbf{S}_\ell^\top\mathbf{Q}_{\ell\ell}^{-1}\mathbf{Q}_{\ell y} \tag{A.16}$$

and

$$= \sum_{\ell=1}^{L}\hat{\mathbf{w}}_\ell. \tag{A.17}$$

Using the above and the fact that $c_\ell^{(j)} = \binom{L-\ell}{j}$, Eq. A.8 can be restated again as

$$\hat{\boldsymbol{\beta}} = \sum_{\ell=1}^{L}\sum_{j=0}^{L-\ell}(-1)^j\binom{L-\ell}{j}\hat{\mathbf{w}}_\ell. \tag{A.18}$$

Finally, Aupetit (2009) shows that for binomial coefficients it holds that

$$\sum_{j=0}^{n}(-1)^j\binom{n}{j} = 0 \quad \forall\, n > 0. \tag{A.19}$$

Eq. A.19 ensures that all values of $\ell < L$ in Eq. A.18 collapse to zero, such that

$$\hat{\boldsymbol{\beta}} = \binom{0}{0}\hat{\mathbf{w}}_L, \tag{A.20}$$

$$= \hat{\mathbf{w}}_L \tag{A.21}$$

and

$$= \mathbf{S}_L^\top\mathbf{Q}_{LL}^{-1}\mathbf{Q}_{Ly}. \tag{A.22}$$

Given that $\mathbf{S}_L$ is the identity matrix, and $\mathbf{z}_L = \mathbf{x}$, this simply becomes the ordinary least squares solution, with

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top y. \tag{A.23}$$

Appendix B. Proof of Proposition 2

Let ν_{eff} be the effective model degrees of freedom of the HFR estimator, with

$$\nu_{\text{eff}} = \text{tr}(\mathbf{P}_{\text{hfr}}) \quad \text{and} \quad \mathbf{P}_{\text{hfr}} = \mathbf{z}\boldsymbol{\Theta}\mathbf{Q}_{zz}^{-1}\mathbf{z}^{\top}.$$

Using Eq. 3.4, the HFR model fit can be written as

$$\hat{y} = \sum_{\ell=1}^{L} \theta_\ell \mathbf{x}\hat{\mathbf{b}}_\ell, \tag{B.1}$$

$$= \sum_{\ell=1}^{L} \theta_\ell \mathbf{x}\mathbf{S}^{\top}(\mathbf{z}_\ell^{\top}\mathbf{z}_\ell)^{-1}\mathbf{z}_\ell^{\top}\mathbf{M}_{\ell-1}y \tag{B.2}$$

and

$$= \sum_{\ell=1}^{L} \theta_\ell \mathbf{P}_\ell \mathbf{M}_{\ell-1}y. \tag{B.3}$$

The projection matrix of the HFR estimator can now be rewritten as

$$\mathbf{P}_{\text{hfr}} = \sum_{\ell=1}^{L} \theta_\ell \mathbf{P}_\ell \mathbf{M}_{\ell-1}, \tag{B.4}$$

$$= \sum_{\ell=1}^{L} \theta_\ell \mathbf{P}_\ell (\mathbf{I}_N - \mathbf{P}_{\ell-1}) \tag{B.5}$$

and

$$= \sum_{\ell=1}^{L} \theta_\ell (\mathbf{P}_\ell - \mathbf{P}_\ell \mathbf{P}_{\ell-1}). \tag{B.6}$$

Here $\mathbf{M}_0 = \mathbf{I}_N$ and $\mathbf{P}_0 = \mathbf{0}$. Recall that for the nested case, where each level contains strictly more information than the preceding level, $\mathbf{M}_{\ell-1} \equiv \prod_{i=1}^{\ell} \mathbf{M}_{\ell-i}$. This implies that $\mathbf{M}_\ell \mathbf{M}_{\ell-1} = \mathbf{M}_\ell$, and by expanding the equality, $\mathbf{P}_\ell \mathbf{P}_{\ell-1} = \mathbf{P}_{\ell-1}$.

Substituting and using the properties of the trace operator, the effective degrees of freedom becomes

$$\text{tr}(\mathbf{P}_{\text{hfr}}) = \sum_{\ell=1}^{L} \theta_\ell \big[\text{tr}(\mathbf{P}_\ell) - \text{tr}(\mathbf{P}_{\ell-1})\big]. \tag{B.7}$$

With a total of $L = K$ levels, the number of features contained in the ℓth level — and thus the rank of $\mathbf{P}_\ell$ — is simply ℓ. The above therefore simplifies to

$$\text{tr}(\mathbf{P}_{\text{hfr}}) = \sum_{\ell=1}^{L} \theta_\ell \big[\ell - (\ell - 1)\big], \tag{B.8}$$

$$= \sum_{\ell=1}^{L} \theta_\ell \cdot 1, \tag{B.9}$$

$$= \sum_{\ell=1}^{L} \theta_\ell. \tag{B.10}$$

Appendix C. Derivation of path-independent HFR estimates

Section 3.5 suggests that $\hat{\boldsymbol{\beta}}_{\text{hfr}}$ can be reformulated to remove path-dependence from the level-specific estimates, with

$$\hat{\boldsymbol{\beta}}_{\text{hfr}} = \hat{\mathcal{B}}\boldsymbol{\phi}. \tag{C.1}$$

To derive this result, recall once again the definition of $\hat{\boldsymbol{\beta}}_{\text{hfr}}$ from Eq. 3.5, and note that using the notation in Appendix A, $\mathbf{S}^\top\boldsymbol{\Theta} = \mathbf{S}_{\boldsymbol{\theta}}^\top$ (i.e. $\mathbf{c}^{(0)} = \boldsymbol{\theta}$), resulting in:

$$\hat{\boldsymbol{\beta}}_{\text{hfr}} = \mathbf{S}^\top\boldsymbol{\Theta}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zy} \tag{C.2}$$

and

$$= \mathbf{S}_{\boldsymbol{\theta}}^\top\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zy}. \tag{C.3}$$

Recall, furthermore, from Appendix A that

$$\hat{\boldsymbol{\beta}} = \sum_{\ell=1}^{L}\sum_{j=0}^{L-\ell}(-1)^j c_\ell^{(j)}\hat{\mathbf{w}}_\ell, \tag{C.4}$$

where $c_\ell^{(j)}$ is no longer equal to the binomial coefficient given that $\mathbf{c}^{(0)} = \boldsymbol{\theta}$.

The definition in Appendix A of $c_\ell^{(j+1)} = \sum_{i=\ell+1}^{L} c_i^{(j)}\ \forall\ \ell < L$ ensures that $c_\ell^{(j+1)} - c_{\ell+1}^{(j)} = c_{\ell+1}^{(j+1)}$. Proceeding inductively, for any arbitrary level $\ell < L$, the appropriate sum in Eq. C.4 can thus be written as:

$$\left(c_\ell^{(0)} - c_\ell^{(1)} + c_\ell^{(2)} - c_\ell^{(3)} + \ldots + (-1)^{L-\ell}c_\ell^{(L-\ell)}\right)\hat{\mathbf{w}}_\ell \tag{C.5}$$

$$= \left(c_\ell^{(0)} - c_{\ell+1}^{(0)} + c_{\ell+1}^{(0)} - c_\ell^{(1)} + c_\ell^{(2)} - c_\ell^{(3)} + \ldots + (-1)^{L-\ell}c_\ell^{(L-\ell)}\right)\hat{\mathbf{w}}_\ell, \tag{C.6}$$

$$= \left(c_\ell^{(0)} - c_{\ell+1}^{(0)} - c_{\ell+1}^{(1)} + c_\ell^{(2)} - c_\ell^{(3)} + \ldots + (-1)^{L-\ell}c_\ell^{(L-\ell)}\right)\hat{\mathbf{w}}_\ell, \tag{C.7}$$

$$= \left(c_\ell^{(0)} - c_{\ell+1}^{(0)} + c_{\ell+1}^{(2)} - c_\ell^{(3)} + \ldots + (-1)^{L-\ell}c_\ell^{(L-\ell)}\right)\hat{\mathbf{w}}_\ell, \tag{C.8}$$

$$= \left(c_\ell^{(0)} - c_{\ell+1}^{(0)} - c_{\ell+1}^{(3)} + \ldots + (-1)^{L-\ell}c_\ell^{(L-\ell)}\right)\hat{\mathbf{w}}_\ell, \tag{C.9}$$

$$\ldots \tag{C.10}$$

$$= \left(c_\ell^{(0)} - c_{\ell+1}^{(0)}\right)\hat{\mathbf{w}}_\ell \tag{C.11}$$

and

$$= \left(\theta_\ell - \theta_{\ell+1}\right)\hat{\mathbf{w}}_\ell. \tag{C.12}$$

When $\ell = L$, $c_\ell^{(j+1)} = 0$, leaving only $\theta_L\hat{\mathbf{w}}_L$. Thus

$$\hat{\boldsymbol{\beta}}_{\text{hfr}} = \sum_{\ell=1}^{L-1}(\theta_\ell - \theta_{\ell+1})\hat{\mathbf{w}}_\ell + \theta_L\hat{\mathbf{w}}_L \tag{C.13}$$

and

$$= \hat{\mathcal{B}}\boldsymbol{\phi}, \tag{C.14}$$

where

$$\boldsymbol{\phi} = \begin{cases} \theta_\ell - \theta_{\ell+1} & \text{when } \ell < L \\ \theta_\ell & \text{otherwise.} \end{cases}$$

Appendix D. Description of growth determinants data set

Table Appendix D.1
Description of growth determinants included in the data set of Sala-I-Martin et al. (2004).

Description	Name	Description	Name
Absolute Latitude	ABSLATIT	Fraction of Land Area Near Navigable Water	LT100CR
Air Distance to Big Cities	AIRDIST	Malaria Prevalence in 1960s	MALFAL66
Ethnolinguistic Fractionalization	AVELF	Fraction GDP in Mining	MINING
British Colony Dummy	BRIT	Fraction Muslim	MUSLIM00
Fraction Buddhist	BUDDHA	Timing of Independence	NEWSTATE
Fraction Catholic	CATH00	Oil Producing Country Dummy	OIL
Civil Liberties	CIV72	Openess measure 1965-74	OPENDEC1
Colony Dummy	COLONY	Fraction Othodox	ORTH00
Fraction Confucian	CONFUC	Fraction Speaking Foreign Language	OTHFRAC
Population Density 1960	DENS60	Primary Schooling in 1960	P60
Population Density Coastal in 1960s	DENS65C	Average Inflation 1960-90	PI6090
Interior Density	DENS65I	Square of Inflation 1960-90	SQPI6090
Population Growth Rate 1960-90	DPOP6090	Political Rights	PRIGHTS
East Asian Dummy	EAST	Fraction Population Less than 15	POP1560
Capitalism	ECORG	Population in 1960	POP60
English Speaking Population	ENGFRAC	Fraction Population Over 65	POP6560
European Dummy	EUROPE	Primary Exports 1970	PRIEXP70
Fertility in 1960s	FERTLDC1	Fraction Protestants	PROT00
Defense Spending Share	GDE1	Real Exchange Rate Distortions	RERD
GDP in 1960 (log)	GDPCH60L	Revolutions and Coups	REVCOUP
Public Education Spending Share in GDP in 1960s	GEEREC1	African Dummy	SAFRICA
Public Investment Share	GGCFD3	Outward Orientation	SCOUT
Nominal Govertnment GDP Share 1960s	GOVNOM1	Size of Economy	SIZE60
Government Share of GDP in 1960s	GOVSH61	Socialist Dummy	SOCIALIST
Gov. Consumption Share 1960s	GVR61	Spanish Colony	SPAIN
Higher Education 1960	H60	Terms of Trade Growth in 1960s	TOT1DEC1
Religion Measure	HERF00	Terms of Trade Ranking	TOTIND
Fraction Hindus	HINDU00	Fraction of Tropical Area	TROPICAR
Investment Price	IPRICE1	Fraction Population In Tropics	TROPPOP
Latin American Dummy	LAAM	Fraction Spent in War 1960-90	WARTIME
Land Area	LANDAREA	War Particpation 1960-90	WARTORN
Landlocked Country Dummy	LANDLOCK	Years Open 1950-94	YRSOPEN
Hydrocarbon Deposits in 1993	LHCPC	Tropical Climate Zone	ZTROPICS
Life Expectancy in 1960	LIFE060		

References

Allaire, J., Tang, Y., 2022. tensorflow: R Interface to 'TensorFlow'. Technical Report R package version 2.9.0.
Aupetit, M., 2009. Nearly Homogeneous Multi-Partitioning with a Deterministic Generator. Neurocomputing 72 (7-9), 1379–1389.
Bach, F., Jenatton, R., Mairal, J., Obozinski, G., 2012. Structured Sparsity through Convex Optimization. Statistical Science 27 (4), 450–468.
Bair, E., Hastie, T., Paul, D., Tibshirani, R., 2006. Prediction by Supervised Principal Components. Journal of the American Statistical Association 101 (473), 119–137.
Bondell, H.D., Reich, B.J., 2008. Simultaneous Regression Shrinkage, Variable Selection and Clustering of Predictors with OSCAR. Biometrics 64(1), 115–123.
Burnham, K.P., Anderson, D.R., 2004. Multimodel Inference — Understanding AIC and BIC in Model Selection. Sociological Methods & Research 33 (2), 261–304.
Bühlmann, P., Rütimann, P., van de Geer, S., Zhang, C.-H., 2013. Correlated Variables in Regression: Clustering and Sparse Estimation. Journal of Statistical Planning and Inference 143 (11), 1835–1858.
Campbell, F., Allen, G.I., 2017. Within Group Variable Selection Through the Exclusive Lasso. Electronic Journal of Statistics 11 (2).
Chen, L., Chen, J., 2015. glmgraph: Graph-Constrained Regularization for Sparse Generalized Linear Models. Technical Report R package version 1.0.3.
Daye, J.Z., Jeng, J.X., 2009. Shrinkage and Model Selection with Correlated Variables Via Weighted Fusion. Computational Statistics & Data Analysis 53 (4), 1284–1298.
Diebold, F.X., Yilmaz, K., 2015. Measuring the Dynamics of Global Business Cycle Connectedness. In: Koopman, S.J., Shephard, N. (Eds.), Unobserved Components and Time Series Econometrics. Oxford University Press, pp. 45–70.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least Angle Regression. Annals of Statistics 32 (2), 407–499.
Eicher, T.S., Papageorgiou, C., Raftery, A.E., 2011. Default Priors and Predictive Performance in Bayesian Model Averaging, with Application to Growth Determinants. Journal of Applied Econometrics 26 (1), 30–55.
Epshtein, B., Uliman, S., 2005. Feature Hierarchies for Object Classification. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. IEEE, Beijing, China, pp. 220–227Vol. 1.
Everitt, B., Landau, S., Stahl, D., Leese, M., 2011. Cluster Analysis, 5th ed Wiley, Chichester, West Sussex, U.K. OCLC: ocn666867900
Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning, Vol. 1, 1 Springer series in statistics Springer, Berlin.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33 (1).
Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Columbus, OH, USA, pp. 580–587.
Grimonprez, Q., Blanck, S., Celisse, A., Marot, G., 2022. MLGL: An R Package Implementing Correlated Variable Selection by Hierarchical Clustering and Group-Lasso. Journal of Statistical Software.

Hansen, B.E., 2007. Least Squares Model Averaging. Econometrica 75 (4), 1175–1189.
Hoerl, A.E., 1962. Application of Ridge Analysis to Regression Problems. Chemical Engineering Progress 58 (3), 54–59.
Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12 (1), 55–67.
Hofmarcher, P., Cuaresma, J.C., Grun, B., Hornik, K., 2011. Fishing Economic Growth Determinants Using Bayesian Elastic Nets. Research Report Series. Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien.
Huang, J., Ma, S., Li, H., Zhang, C.-H., 2011. The Sparse Laplacian Shrinkage Estimator for High-Dimensional Regression. The Annals of Statistics 39 (4).
Huang, J., Zhang, T., 2010. The Benefit of Group Sparsity. Annals of Statistics 38, 1978–2004.
Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., Shang, H.L., 2011. Optimal Combination Forecasts for Hierarchical Time Series. Computational Statistics & Data Analysis 55 (9), 2579–2589.
Jacob, L., Obozinski, G., Vert, J.-P., 2009. Group Lasso with Overlap and Graph Lasso. In: Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09. ACM Press, Montreal, Quebec, Canada, pp. 1–8.
James, W., Stein, C., 1961. Estimation with Quadratic Loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1, 361–380.
Jolliffe, I.T., 2002. Principal Component Analysis, 2 Springer, New York.
Kaufman, L., Rousseeuw, P.J., 2005. Finding Groups in Data: An Introduction to Cluster Analysis, 1 Wiley, Hoboken, N.J.
Kim, S., Xing, E.P., 2012. Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity. The Annals of Applied Statistics 6 (3), 1095–1117.
Kose, M.A., Otrok, C., Whiteman, C.H., 2003. International Business Cycles: World, Region, and Country-Specific Factors. The American Economic Review 93 (4).
Ledoit, O., Wolf, M., 2003. Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection. Journal of Empirical Finance 10 (5), 603–621.
Ley, E., 2008. On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression. MPRA Papers.
Li, C., Li, H., 2008. Network-Constrained Regularization and Variable Selection for Analysis of Genomic Data. Bioinformatics 24 (9), 1175–1182.
Li, C., Li, H., 2010. Variable Selection and Regression Analysis for Graph-Structured Covariates with an Application to Genomics. The Annals of Applied Statistics 4 (3).
Maechler, M., Rousseeuw, P., Struyf, A., Hornik, K., 2019. cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.
Maimon, O., Rokach, L., 2010. Data Mining and Knowledge Discovery Handbook, 2 Springer US, Boston, MA.
Mallows, C., 1973. Some Comments on CP. Technometrics 15 (4), 661–675.
Martens, H., 2001. Reliable and Relevant Modelling of Real World Data: A Personal Account of the Development of PLS Regression. Chemometrics and Intelligent Laboratory Systems 58 (2), 85–95.
Mevik, B.-H., Wehrens, R., 2019. Introduction to the pls Package. R package manuals.
Mishra, C., Gupta, D.L., 2017. Deep Machine Learning and Neural Networks: An Overview. IAES International Journal of Artificial Intelligence (IJ-AI) 6 (2).
Pfitzinger, J., 2023. hfr: Estimate Hierarchical Feature Regression Models. Technical Report R package version 0.6.2.
Qiu, L., Qu, Y., Shang, C., Yang, L., Chao, F., Shen, Q., 2021. Exclusive Lasso-Based K-Nearest-Neighbor Classification. Neural Computing and Applications 33 (21), 14247–14261.
Rey, H., 2015. Dilemma not Trilemma: The Global Financial Cycle and Monetary Policy Independence. NBER Working Papers No. 21162.
Roth, V., Fischer, B., 2008. The Group-Lasso for Generalized Linear Models: Uniqueness of Solutions and Efficient Algorithms. Proceedings of the International Conference on Machine Learning (ICML).
Sala-I-Martin, X., Doppelhofer, G., Miller, R.I., 2004. Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach. The American Economic Review 94 (4).
Schneider, U., Wagner, M., 2012. Catching Growth Determinants with the Adaptive Lasso: Lassoing Growth Determinants. German Economic Review 13 (1), 71–85.
Schäfer, J., Strimmer, K., 2005. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. Statistical Applications in Genetics and Molecular Biology 4 (1).
Sharma, D.B., Bondell, H.D., Zhang, H.H., 2013. Consistent Group Identification and Variable Selection in Regression With Correlated Predictors. Journal of Computational and Graphical Statistics 22 (2), 319–340.
Shen, X., Huang, H.-C., Pan, W., 2012. Simultaneous Supervised Clustering and Feature Selection Over a Graph. Biometrika 99 (4), 899–914.
Simon, H., 1962. The Architecture of Complexity. Proceedings of the American Philosophical Society 106 (6), 467–482.
Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A Sparse-Group Lasso. Journal of Computational and Graphical Statistics 22 (2), 231–245.
Stock, Watson, 2016. Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. In: Handbook of Macroeconomics, Vol. 2. Elsevier, pp. 415–525.
Stock, Watson, 2016. Factor Models and Structural Vector Autoregressions in Macroeconomics. Handbook of Macroeconomics 2.
Szafranski, M., Grandvalet, Y., Morizet-Mahoudeaux, P., 2007. Hierarchical Penalization. Advances in Neural Information Processing Systems 20, 1457–1464.
Taylor, A., Tibshirani, R., 2022. genlasso: Path Algorithm for Generalized Lasso Problems. Technical Report R package version 1.6.1.
Core Team, R., 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.
Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58 (1), 267–288.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and Smoothness via the Fused Lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (1), 91–108.
Turlach, B.A., Venables, W.N., Wright, S.J., 2005. Simultaneous Variable Selection. Technometrics 47 (3), 349–363.
Ulbricht, J., 2012. lqa: Penalized Likelihood Inference for GLMs. Technical Report R package version 1.0-3.
Varian, H.R., 2014. Big Data: New Tricks for Econometrics. Journal of Economic Perspectives 28 (2), 3–28.
Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58 (301), 236–244.
Weylandt, M., Campbell, F., 2018. ExclusiveLasso: Generalized Linear Models with the Exclusive Lasso Penalty. Technical Report R package version 0.0.
Witten, D.M., Shojaie, A., Zhang, F., 2014. The Cluster Elastic Net for High-Dimensional Regression With Unknown Variable Grouping. Technometrics 56 (1), 112–122.
Wold, S., 2001. Personal Memories of the Early PLS Development. Chemometrics and Intelligent Laboratory Systems 58 (2), 83–84.
Yang, Y., Zou, H., Bhatnagar, S., 2020. gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm. Technical Report R package version 1.5.
Yuan, M., Lin, Y., 2006. Model Selection and Estimation in Regression with Grouped Variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1), 49–67.
Zeng, X., Figueiredo, M.A.T., 2013. A Novel Sparsity and Clustering Regularization. Paper. arXiv.org.
Zhao, P., Rocha, G., Yu, B., 2009. The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection. The Annals of Statistics 37 (6A).
Zhou, Y., Jin, R., Hoi, S.C.H., 2010. Exclusive Lasso for Multi-task Feature Selection. International Conference on Artificial Intelligence and Statistics 988–995.
Zou, H., 2006. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association 101 (476), 1418–1429.
Zou, H., Hastie, T., 2005. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2), 301–320.
Zou, H., Zhang, H.H., 2009. On the Adaptive Elastic-Net with a Diverging Number of Parameters. The Annals of Statistics 37 (4), 1733–1751.