“On the “mementum” of meme stocks”
Online Appendix

Michele Costola* Matteo Iacopini† Carlo R.M.A. Santagiustina‡

A Econometric Model

The time-varying rank TVP-VECM model of Chua and Tsiaplias (2018) assumes two sources of temporal variation in the cointegrating matrix of a VECM, Π. First, they consider the SVD decomposition of Π

$$\Pi = U\Lambda V',$$

where $U'U = I_n$, $V'V = I_n$, and Λ is an n-dimensional diagonal matrix with diagonal elements $w_1 \geq \ldots \geq w_n \geq 0$. In presence of $r < n$ cointegrating relationships, the matrix Π has rank r, meaning that last $n - r$ singular values of Π are zero. Moreover, it admits the low rank decomposition

$$\begin{aligned}
\Pi &= U\Lambda V' = \beta\alpha = U_{1:r}\Lambda_{1:r,1:r}V'_{1:r} \\
\alpha &= \Lambda_{1:r,1:r}V'_{1:r} \\
\beta &= U_{1:r},
\end{aligned}$$

where $U_{1:r}$ denote the first r columns of U (similarly for V) and $\Lambda_{1:r,1:r}$ represent the top-left square submatrix of Λ of size r.

To account for time-varying cointegration rank, Chua and Tsiaplias (2018) introduce an idempotent diagonal matrix into the SVD decomposition of Π, denoted $I(S_t)$, and assume it is driven by a hidden homogeneous finite state Markov chain, S_t. This results in

$$\Pi_t = UI(S_t)I(S_t)\Lambda V' = UI(S_t)I(S_t)D, \tag{1}$$

*Ca' Foscari University of Venice and Leibniz Institute for Financial Research SAFE, Email: michele.costola@unive.it.

†Vrije Universiteit Amsterdam and Tinbergen Institute, Email: m.iacopini@vu.nl.

‡Ca' Foscari University of Venice and Venice International University, Email: carlo.santagiustina@unive.it.

with $D = \Lambda V'$. The matrix $I(S_t)$ has generic k-th diagonal element

$$I(S_t)_{kk} = (1 - s_{1t}) \sum_{j=k+1}^{n+1} s_{jt}, \qquad k = 1, \dots, n, \tag{2}$$

where

$$s_{jt} = \begin{cases} 1 & \text{if } S_t = j \\ 0 & \text{otherwise.} \end{cases}$$

From eq. (2) it follows that when $s_{jt} = 1$ the rank of Π_t is exactly $r = (j-1)$, thus providing a direct link between the state of the hidden Markov chain, S_t, and the number of cointegrating relationships, r. Considering an n-dimensional system, the rank ranges from $r = 0$ (resulting in a VAR model for δy_t) to $r = n$ (meaning a VAR model for stationary y_t). Therefore, the homogeneous Markov chain has $n + 1$ possible states, from $j = 1$ to $j = n + 1$, and the transition probabilities are encoded in the $(n + 1 \times n + 1)$ matrix P whose generic entry is given by $P_{ij} = \mathbb{P}(S_t = j | S_{t-1} = 1)$, for $i, j = 1, \dots, n + 1$.

Finally, to allow for smooth changes over time of both the loading matrix, α, and the cointegrating relationships, β, Chua and Tsiaplias (2018) assume that matrices U and V from the SVD decomposition of Π are time-varying. To do that, one must first of all address the identifiability issue of the VECM. In particular, since the factorization $\Pi = \beta\alpha$ does not allow to separately identify all the entries of β and α, a parameter expansion technique is used, following Koop et al. (2011), leading to

$$\begin{aligned} \Pi_t &= U_t I(S_t) I(S_t) D_t \\ &= U_t \kappa_t I(S_t) I(S_t) \kappa_t^{-1} D_t \\ &= U_t^* I(S_t) I(S_t) D_t^* \\ &= \beta^* \alpha_t^*, \end{aligned} \tag{3}$$

where $\kappa_t = (U_t^{*\prime} U_t^*)^{1/2}$, $U_t^* = U_t \kappa_t$, $D_t^* = \kappa_t^{-1} D_t$, and

$$\alpha_t^* = I(S_t) \kappa_t^{-1} D_t = I(S_t) D_t^* \tag{4}$$

$$\alpha_t^* = U_t \kappa_t I(S_t) = U_t^* I(S_t). \tag{5}$$

However, it is important to recall that, as both U_t^* and D_t^* stem from the SVD decomposition of Π_t, they must satisfy an orthogonality condition. In particular, one has to ensure that $(U_t^{*\prime} U_t^*) = I_n$ and $(D_t^* D_t^{*\prime}) = \Lambda\Lambda$. To ensure these conditions hold, Chua and Tsiaplias (2018) impose $n(n-1)/2$ constraints on U_t^*, by restricting the first $j - 1$ elements of the

j-th column, for $j = 2, \ldots, n$, and $n(n-1)/2$ constraints on D_t^*, by restricting the first $j-1$ elements of the j-th row, for $j = 2, \ldots, n$. Overall, this results in $n^* = n(n+1)/2$ unrestricted parameters located in the lower triangular part of U_t^* and n^* unrestricted parameters located in the upper triangular part of D_t^*.

Finally, defining $u_t^* = \text{vec}(U_t^*)$ and $d_t^* = \text{vec}(D_t^*)$, the smooth evolution of the loadings is modelled by assuming

$$d_{t+1}^* = d_t * + \zeta_t, \qquad \zeta_t \sim \mathcal{N}(0, Q), \tag{6}$$

where $Q = \text{diag}(\sigma_1^2, \ldots, \sigma_{n^*}^2)$ and $d_0^* \sim \mathcal{N}(0, \sigma_d I_{n^*})$. Instead the dynamics of the cointegrating relationships is defined by the stationary VAR process

$$u_{t+1}^* = \rho u_t^* + \eta_t, \qquad \mathcal{N}(0, I_{n^*}), \tag{7}$$

with $|\rho| < 1$ and $u_0^* \sim \mathcal{N}(0, 1/(1-\rho^2))$.

A.1 Prior Distributions and Posterior Inference

We adopt the same prior specification as in Chua and Tsiaplias (2018), that is we assume

$$\begin{aligned}
\mathbf{g} &\sim \mathcal{N}(\mathbf{g}|\underline{\mu}_g, \underline{\Sigma}_g) \\
\Sigma &\sim \mathcal{IW}(\Sigma|\underline{\nu}, \underline{\Psi}) \\
P_i &\sim \mathcal{D}ir(P_i|\underline{\tau}_{i,1}, \ldots, \underline{\tau}_{i,n+1}), \quad i = 1, \ldots, n+1 \\
\sigma_i^2 &\sim \mathcal{IG}\Big(\sigma_i^2\Big|\frac{\underline{\theta}_i}{2}, \frac{\underline{\theta}_i \underline{f}_i}{2}\Big), \quad i = 1, \ldots, n^* \\
\rho &\sim \mathcal{U}(\rho|\underline{\rho}_0, \underline{\rho}_1),
\end{aligned}$$

where P_i is the i-th row of the transition matrix P.

To describe the MCMC algorithm, first define $\mathbf{S} = (S_1, \ldots, S_T)'$, $\mathbf{y} = (y_1, \ldots, y_T)$, $\mathbf{d}^* = (d_1^*, \ldots, d_T^*)$, $\mathbf{u}^* = (u_1^*, \ldots, u_T^*)$, $\mathbf{g} = \text{vec}((c', B')')$. The most involved steps of the MCMC are those concerned with the sampling of the time-varying parameters, $\mathbf{d}^*$ and $\mathbf{u}^*$, which represent the unrestricted elements of the loading matrix and cointegrating relationships, respectively. Conditioning on the observations and the other parameters, one obtains a linear Gaussian state space model with d_t^* as latent state vector. Therefore, it is possible to sample the path $\mathbf{d}^*$ by relaying on Kalman filter and smoother techniques. An analogous approach can be used to sample the path $\mathbf{u}^*$. Instead, the path of the hidden states, $\mathbf{S}$, is drawn using a multi-move Gibbs sampler (Frühwirth-Schnatter, 2006; Kim et al., 1998; Carter and Kohn, 1994).

Concerning the constant parameters, conditionally on the observations and all the time varying parameters, the observation model can be expresses as a linear Gaussian SUR, thus allowing for the use of conjugate priors for the coefficients, $\mathbf{g}$, and the covariance matrix, Σ. Similarly, we can easily draw σ_i^2 from its full conditional, as it admits a conjugate inverse gamma prior distribution. Instead, the autoregressive coefficient driving the dynamics of the cointegrating vectors, ρ, is sampled using a Metropolis-Hastings step as in Koop et al. (2011).

Summarizing, the Gibbs sampler cycles over the following steps:

1. sample $\mathbf{u}^*$, conditionally on $\mathbf{y}, \mathbf{S}, \mathbf{d}^*, \mathbf{g}, \rho, \Sigma$, using the Kalman filter and smoother with the simulation smoother method of Durbin and Koopman (2012).

2. sample the latent variables autoregressive coefficient ρ, conditionally on $\mathbf{u}^*$, from its full conditional distribution using a Metropolis-Hastings step as in Koop et al. (2011).

3. sample $\mathbf{d}^*$, conditionally on $\mathbf{y}, \mathbf{S}, \mathbf{u}^*, \mathbf{g}, \sigma, \Sigma$, using the Kalman filter and smoother with the simulation smoother method of Durbin and Koopman (2012).

4. sample the latent variables variance parameter, σ_i^2, for $i = 1, \ldots, n$, conditionally on $\mathbf{d}^*$, from its full conditional distribution $p(\sigma_i^2 | d_{i,1}^*, \ldots, d_{i,T}^*) = \mathcal{IG}(\sigma_i^2 | \overline{\theta}_i, \overline{f}_i)$.

5. sample the coefficients of the VECM, $\mathbf{g}$, conditionally on $\mathbf{y}, \mathbf{S}, \mathbf{d}^*, \mathbf{u}^*, \Sigma$ from its full conditional distribution $p(\mathbf{g} | \mathbf{y}, \mathbf{S}, \mathbf{d}^*, \mathbf{u}^*, \Sigma) = \mathcal{N}(\mathbf{g} | \overline{\mu}_g, \overline{\Sigma}_g)$.

6. sample the covariance matrix, Σ, conditionally on $\mathbf{y}, \mathbf{S}, \mathbf{d}^*, \mathbf{u}^*, \mathbf{g}$ from its full conditional distribution $p(\Sigma | \mathbf{y}, \mathbf{S}, \mathbf{d}^*, \mathbf{u}^*, \mathbf{g}) = \mathcal{IW}(\Sigma | \overline{\nu}, \overline{\Psi})$.

7. sample the latent states S_t, $t = 1, \ldots, T$, conditionally on $\mathbf{y}, \mathbf{d}^*, \mathbf{u}^*, \mathbf{g}, P, \Sigma$, using a multi-move Gibbs Sampling algorithm (Frühwirth-Schnatter, 2006; Kim et al., 1998; Carter and Kohn, 1994).

8. sample the rows of the transition matrix P_i, $i = 1, \ldots, n+1$, from its full conditional distribution $p(P_i | \mathbf{S}) = \mathcal{D}ir(P_i | \overline{\tau})$.

B Data collection

We consider the following six stocks: GameStop (GME), AMC Entertainment (AMC), KOSS Corporation (KOSS), Moody's (MCO), Pfizer (PFE), and Disney (DIS). The considered period is from January 2019 to April 2021 at daily frequency. Stock prices and stock trading volumes have been downloaded in Bloomberg. Tweets related to the aforementioned stocks have been collected through a R script based on the *academictwitter* library (Barrie and ting Ho, 2021), which programmatically sends requests to Twitter's FullArchive API v2 (see: `https://developer.twitter.com/en/docs/twitter-api/early-access`). By repetitively querying the Twitter's API we have collected Twitter posts published from January 2019 to April 2021 that jointly match the following conditions: (i) contain one or more images (ii) are not a retweet, and (iii) contain one or more Twitter hash-tag (#[ACRONYM]) or fin-tag[1] ($[ACRONYM]) explicitly referring to the targeted stock. The specific query conditions for each stock are reported in Table 1.

Stock name (acronym)	**condition (i)**	**condition (ii)**	**condition (iii)**
GameStop (GME)	`-is:retweet`	`has:images`	`#GME OR $GME`
AMC Entertainment (AMC)			`#AMC OR $AMC`
KOSS Corporation (KOSS)			`#KOSS OR $KOSS`
Moody's (MCO)			`#MCO OR $MCO`
Pfizer (PFE)			`#PFE OR $PFE`
Disney (DIS)			`#DIS OR $DIS`

Table 1: List of query conditions used for retrieving Twitter data from API V2. for each stock, all conditions must be met to include a Twitter post in our data-set.

For each stock, raw tweets matching the three conditions are then transformed in count time series at the daily frequency. Only working days for which market data is also available are considered in the empirical analysis.

References

Barrie, C. and ting Ho, J. C. (2021). academictwitter: an r package to access the twitter academic research product track v2 api endpoint. *Journal of Open Source Software*, 6(62):3272.

[1]fin-tags ($) are used in Twitter to talk about a publicly-traded company.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Chua, C. L. and Tsiaplias, S. (2018). A Bayesian approach to modeling time-varying cointegration and cointegrating rank. *Journal of Business & Economic Statistics*, 36(2):267–277.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford university press.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Kim, C.-J., Nelson, C. R., and Startz, R. (1998). Testing for mean reversion in heteroskedastic data based on gibbs-sampling-augmented randomization. *Journal of Empirical finance*, 5(2):131–154.

Koop, G., Leon-Gonzalez, R., and Strachan, R. W. (2011). Bayesian inference in a time varying cointegration model. *Journal of Econometrics*, 165(2):210–220.