A Regularized Vector Autoregressive Hidden Semi-Markov model, with Application to Multivariate Financial Data

Zekun Xu
North Carolina State University
Raleigh, NC USA
zxu13@ncsu.edu

Ye Liu
North Carolina State University
Raleigh, NC USA
yliu87@ncsu.edu

Abstract

Hidden Markov model (HMM) has been a popular choice for financial time series modeling due to its advantage in capturing dynamic regimes. However, HMM's implicit assumption that the state duration follows a geometric distribution is too strong to hold in practice. In this work, we propose a regularized vector autoregressive hidden semi-Markov model to analyze multivariate financial time series. One challenge in such a model setting is that the number of parameters is too large to be reliably estimated unless the time series is extremely long. To address this issue, an augmented EM algorithm is developed for parameter estimation by using regularized estimators for the state-dependent covariance matrices and autoregression matrices in the M-step. The performance of the proposed model is evaluated in a simulation experiment, and demonstrated with the New York Stock Exchange financial portfolio data.

Introduction

In finance and economics, it is often assumed that the financial returns follow a white noise process. However, empirical evidence suggests that this assumption may be too strong to hold in practice. Ding, Granger, and Engle (1993) found that there is substantial correlation between absolute returns. Andersen et al. (2001) indicated that realized volatilities and correlations show strong temporal dependence and appear to be well described by long-memory processes. Moreover, Fan and Yao (2017) commented that the squared and the absolute returns of S&P 500 index exhibits significant serial correlations. Therefore, it is reasonable to model the financial return series using an autoregressive process.

The drawback of an autoregressive process is that it alone cannot model the volatility clustering and heavy-tailed distribution in the financial return series. This is because such financial return series often have more than one latent data generating mechanisms. For example, the performance of a financial portfolio in a stable economy is expected to follow a different autoregressive process from that in a volatile economy. Rydén, Teräsvirta, and Åsbrink (1998) showed that a hidden Markov model (HMM) can reproduce most of the stylized facts for daily return series Granger and Ding (1995).

HMM is a bivariate discrete time stochastic process $\{S_t, Y_t\}_{t\geq 0}$ such that

(A1) $\{S_t\}$ is a Markov chain, i.e. $P(S_t|S_{t-1}, ..., S_1) = P(S_t|S_{t-1})$.
(A2) $\{Y_t\}$ is a sequence of conditional independent random variables given $\{S_t\}$.

In a Gaussian HMM, the marginal distributions for observed series are essentially modeled as a mixture of Gaussian distributions such that volatility clustering and heavy-tailedness are automatically incorporated in the model framework. Further, the transition between the latent states are directly modeled in HMM so as to account for the temporal dependence in the series.

However, assumptions (A1) and (A2) may both be too strong to hold in practice. Assumption (A1) indicates that the current latent state depends only on the most recent latent state in the past; beyond that, it is memoryless. Rydén, Teräsvirta, and Åsbrink (1998) illustrated that the stylized fact of the very slowly decaying autocorrelation for absolute (or squared) returns cannot be described by a HMM. Bulla and Bulla (2006) proposed the use of hidden semi-Markov model (HSMM) to overcome the lack of flexibility of HMM to model the temporal higher order dependence in financial returns. In HSMM, the latent state durations are explicitly modeled rather than assuming them to be geometric as in HMM. This has the practical advantage since it is typical that the longer time the economy spends in one of the latent states the more likely it will switch to another latent state. In the meantime, assumption (A2) can be dropped in the class of Markov-switching models proposed by Hamilton (1989) where $\{Y_t\}$ is allowed to follow state-dependent Gaussian vector autoregressive processes, also known as vector autoregressive hidden Markov models (VAR-HMM). Yang (2000) pointed out another interesting feature that VAR-HMM can occasionally behave in a nonstationary manner although being stationary and mean reverting in the long run.

For general applicability, we are going to adopt the most flexible framework of a p^{th} order vector autoregressive hidden semi-Markov model [VAR(p)-HSMM] to analyze multivariate financial time series. Note that both VAR and HMM are special cases in the VAR(p)-HSMM framework. Our goal is to make inference on the parameters that determine the data generating mechanism, as well as evaluate the prediction performance. A potential problem of VAR(p)-HSMM is the large

number of parameters to be estimated when the dimension of Y_t is high. A multivariate M-state VAR(p)-HSMM series of dimension d has $\frac{Md(d+1)}{2}$ parameters in the state-dependent covariance matrices and Mpd^2 parameters in the autoregression matrices. Unless the time series is extremely long, we are not able to reliably estimate the covariance and autoregression matrices even when the dimension d is moderate. Therefore, regularizations are needed to stabilize the parameter estimation. Städler and Mukherjee (2013), Fiecas et al. (2017), and Monbet and Ailliot (2017) proposed different versions of a penalized log-likelihood procedure with regularization on the state-dependent inverse covariance matrices in a Gaussian HMM to form a more stable regularized estimator. So far, there is no literature that elaborates on the regularized estimation for VAR(p)-HSMM framework. Neither has the regularized VAR(p)-HSMM framework been used to model multivariate financial returns yet.

Thus, our contribution is to provide a detailed parameter estimation procedure for a regularized VAR(p)-HSMM. The model framework of VAR(p)-HSMM is provided in Section 2, where we integrated the LASSO regularization by Tibshirani (1996) on autoregression matrices, and shrinkage regularization by Ledoit and Wolf (2004) on covariance matrices into the EM algorithm. Section 3 presents simulation studies on finite samples to evaluate the performance of the proposed regularized estimators in different scenarios. Section 4 provides an empirical analysis on the NYSE financial portfolio of 50 stocks using the regularized VAR(p)-HSMM. Section 5 gives a brief discussion. All the analyses utilize the R package "rarhsmm", which has been developed for fitting regularized VAR(p)-HSMM.

Methodology

Model framework

Denote by $\mathbf{y}_t \in \mathbb{R}^d$ for t=1,...,T to be the observed multivariate data at time t, where d is the dimension for each $\mathbf{y}_t$. Denote by $S_t \in \{1, ..., M\}$ to be the latent state at time t, where M is the fixed finite number of states. Let $\boldsymbol{\delta} = [\delta_1, ..., \delta_M]$ be the prior probability of latent states. Further, we denote the latent state duration densities by $\mathbf{r} = [r_1, ..., r_M]$ such that

$$r_i(n) = P(\text{stay n times in latent state i}) \qquad n = 1, 2, ..., D,$$

where D is the fixed maximum state duration, i.e. any state duration greater than D will be censored at D. In addition, denote by $\mathbf{Q} = \{q_{ij}\}$ for i=1,...,M and j=1,...,M the state transition matrix such that

$$q_{ij} = P(S_{t+1} = j | S_t = i) \qquad t = 1, ..., T-1,$$

where $\sum_{j=1}^{M} q_{ij} = 1 \quad \forall \ \ i \in 1, ..., M$

Thus, the data generating mechanism for VAR(p)-HSMM, can be described as follows. First, an initial state, $S_1 = i$ ($i \in 1, ..., M$) is chosen according to the initial state distribution δ_i. Second, a duration n is chosen according to the latent state duration density $r_i(n)$. Third, observations $\mathbf{y}_1, ..., \mathbf{y}_n \in \mathbb{R}^d$ are chosen according to the state-dependent p^{th} order Gaussian vector autoregressive process

$$\mathbf{y}_t = \boldsymbol{\mu}_i + \sum_{k=1}^{p} \mathbf{A}_{ki} \mathbf{y}_{t-k} + \boldsymbol{\epsilon}_{ti} \text{ where } \boldsymbol{\epsilon}_{ti} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i), \qquad (1)$$

for i = 1,...,M and t = 1,...,n, where $\boldsymbol{\mu}_i \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{d \times d}$ are the conditional mean and covariance matrix of $\mathbf{y}_t$ given $S_t, \mathbf{y}_{t-1}, ..., \mathbf{y}_{t-p}$; $\mathbf{A}_{ki} \in \mathbb{R}^{d \times d}$ is the k^{th}-order autoregression matrix conditioning on $S_t = i$.

Fourth, the next state, $S_{n+1} = j$, is chosen according to the state transition probability q_{ij}, the i, j^{th} element in the transition matrix $\mathbf{Q}$. An implicit constraint is that there should be no transition back to the same state because we generate exactly n observations in latent state i in the previous steps, i.e. $S_{1:n} = i$. Then the data generating process repeats the previous steps until we end up with T observations.

Denote by $\boldsymbol{\theta} = [\boldsymbol{\delta}, \mathbf{r}, \mathbf{Q}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}]$ the set of all parameters in VAR(p)-HSMM, where there are $M - 1$ free parameters in $\boldsymbol{\delta}$, $M(D-1)$ in $\mathbf{r}$, $M(M-2)$ in $\mathbf{Q}$, Md in $\boldsymbol{\mu}$, $\frac{Md(d+1)}{2}$ in $\boldsymbol{\Sigma}$, and Mpd^2 in $\mathbf{A}$.

Our VAR(p)-HSMM framework is a natural generalization of the VAR(p)-HMM framework (Hamilton, 1989; Yang, 2000; Monbet and Ailliot, 2017; Francq and Zakoıan, 2001) by allowing for the explicit modeling of the state duration distributions. In particular, we set all the latent state duration densities to be discrete nonparametric distributions with arbitrary point mass assigned to the feasible duration values so as to allow for the most flexibility.

Regularization

There are two motivations for us to apply regularization on the VAR(p)-HSMM framework. On the one hand, the daily financial time series is typically not long enough for us to reliably estimate all the parameters in the state-dependent covariance matrices in the VAR(p)-HSMM. Those covariance matrices may not be invertible especially when the dimension of $\mathbf{y}_t$ is high. On the other hand, we assume that the state-dependent autoregression matrices to be sparse, i.e. many entries are nearly zero. Although the white noise assumption is often used in financial return data, the empirical evidence indicates that the IID assumption is too strong and too restrictive to be true in general (Fan and Yao, 2017; Franke, Härdle, and Hafner, 2004). Thus, a regularized estimator for autoregression matrices can shrink the negligible correlations to zero while allow for the possibility that some correlations may be significant.

The regularized estimator for state-dependent covariance matrices follows the work of Ledoit and Wolf (2004), Sancetta (2008), and Fiecas et al. (2017) such that each regularized estimator is a convex combination of the maximum likelihood estimator and a scaled identity matrix with the same trace,

$$\boldsymbol{\Sigma}^r = \frac{1}{1+\lambda_\Sigma} \hat{\boldsymbol{\Sigma}}^{mle} + \frac{\lambda_\Sigma}{1+\lambda_\Sigma} c\mathbf{I} \quad s.t \quad \text{tr}(\hat{\boldsymbol{\Sigma}}^{mle}) = \text{tr}(c\mathbf{I}),$$

where $\lambda_\Sigma \geq 0$ controls the strength of the regularization. Note that when $\lambda_\Sigma = 0$, we have $\boldsymbol{\Sigma}^r = \hat{\boldsymbol{\Sigma}}^{mle}$. This regularized estimator results in shrinkage on the covariance estimates and ensures the positive definiteness of the estimated covariance matrix when the sample covariance matrix is close to singularity. This holds even if λ_Σ is very small so that we do not increase much bias when stabilizing the estimate. Besides, this regularization yields not only invertible but also

well-conditioned covariance estimates. As λ_Σ increases, the dispersion between the smallest and the largest eigenvalues for the estimated covariance matrix shrinks so that the matrix becomes more regular.

The regularized estimator for state-dependent autoregressive coefficients is based on the classic LASSO regularizationTibshirani (1996) such that

$$\mathbf{a}^r = \arg\min_{\mathbf{a}} \| \text{vec}(\mathbf{y}_{p+1:T}) - \mu + \sum_{k=1}^{p} \mathbf{a}_k^\top \text{vec}(\mathbf{y}_{p+1-k:T-k}) \|_2^2 + \lambda_a \|\mathbf{a}\|_1,$$

where vec is the vectorization operator, and $\mathbf{a} = [\mathbf{a}_p^\top, ..., \mathbf{a}_1^\top]^\top = [\text{vec}(\mathbf{A}_p)^\top, ..., \text{vec}(\mathbf{A}_1)^\top]^\top$ is the vectorization of the state-dependent autoregression matrices. Here $\lambda_a \geq 0$ controls the strength of the regularization on the ℓ_1 LASSO penalty, i.e. a larger λ_a will induce a more sparse estimator.

Cross-validation

The selection of the optimal regularization parameters λ_Σ and λ_a will be based on a similar cross-validation scheme by minimizing one-step-ahead mean-square forecast error (MSFE) as was described in Bańbura, Giannone, and Reichlin (2010) and Nicholson, Matteson, and Bien (2014). More specifically, the data is divided into three periods: one for training (1:T_1), one for validation (T_1:T_2), and one for forecasting (T_2:T).

The validation process starts by fitting a model using all data up to time T_1 and forecast $\mathbf{y}_{T_1+1}^{\lambda_\Sigma,\lambda_a}$. Then we sequentially add one observation at a time and repeat this process until time T_2. Finally, from time T2 to T, we evaluate the one-step-ahead forecast error by minimizing

$$MSFE(\lambda_\Sigma, \lambda_a) = \frac{1}{T_2 - T_1} \sum_{t=T_1}^{T_2-1} \|\mathbf{y}_{t+1}^{\lambda_\Sigma,\lambda_a} - \mathbf{y}_{t+1}\|_F^2,$$

where $\|.\|_F$ is the Frobenius norm defined as $\|\mathbf{A}\|_F = \sqrt{\text{tr}(\mathbf{A}^\top \mathbf{A})}$. A two-dimensional grid-search is adopted to find the regularization values that minimize the MSFE, with 15 grid points in each dimension.

Parameter estimation

The parameter estimation procedure follows the general framework of EM algorithm for the class of hidden Markov models proposed by Baum et al. (1970) and popularized by Dempster, Laird, and Rubin (1977). Regarding the implementation of the EM algorithm to maximize the penalized likelihood function, the monotonic property and convergence results have been proved in Green (1990) and De Pierro (1995).

In the E-step, the standard forward-backward variables are generalized on the basis of Rabiner (1989) and Yu (2010). Define

$$f_{j,n}(\mathbf{y}_{t+1:t+n}) = P(\mathbf{y}_{t+1:t+n} | S_{t+1:t+n} = j),$$

i.e. the state-dependent multivariate autoregressive Gaussian density for state j that lasts for duration n. Then, define the forward variables

$$\alpha_t(j,n) = P(S_{t-n+1:t} = j, \mathbf{y}_{1:t} | \boldsymbol{\theta}),$$

where $j = 1, ..., M$, $t = 1, ..., T$, and $n = \{1, ..., \min(D, t)\}$. Initialize

$$\alpha_0(j,n) = \delta_j \qquad j = 1, ..., M, \tag{2}$$

Define the recursion

$$\alpha_t(j,n) = \sum_{i=1}^{M} \sum_{n'=1}^{\min(D,t)} \alpha_{t-n}(i,n') q_{ij} r_i(n) f_{j,n}(\mathbf{y}_{t-n+1:t}). \tag{3}$$

Similarly, define the backward variables $\beta_t(j,n) = P(\mathbf{y}_{t+1:T} | S_{t-n+1:t} = j, \boldsymbol{\theta})$ where $j = 1, ..., M$, $t = 1, ..., T$, and $n = \{1, ..., \min(D, t)\}$. Initialize $\beta_T(j,n) = 1$ and define the recursion

$$\beta_t(j,n) = \sum_{i=1}^{M} \sum_{n'=1}^{\min(D,T-t)} q_{ji} r_j(n) f_{i,n'}(\mathbf{y}_{t+1:t+n'}) \beta_{t+n'}(i,n'). \tag{4}$$

In addition, define the following 3 sets of auxiliary variables

$$\begin{aligned} \xi_t(i,j) &= P(S_t = i, S_{t+1} = j, \mathbf{y}_{1:T} | \boldsymbol{\theta}) \\ &= \sum_{n'=1}^{\min(D,t)} \sum_{n=1}^{\min(D,T-t)} \alpha_t(i,n') q_{ij} f_{j,n}(\mathbf{y}_{t+1:t+n}) \beta_{t+n}(j,n), \end{aligned} \tag{5}$$

$$\eta_t(j,n) = P(S_{t-n+1:t} = j, \mathbf{y}_{1:T} | \boldsymbol{\theta}) = \alpha_t(j,n) \beta_t(j,n), \tag{6}$$

$$\gamma_t(j) = P(S_t = j, \mathbf{y}_{1:T} | \boldsymbol{\theta}) = \sum_{n=1}^{\min(D,t,T-t)} \eta_t(j,n). \tag{7}$$

Then in the E-step, we are ready to compute

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}) =& E_{\boldsymbol{\theta}^{(l)}} \{\log[P_{\boldsymbol{\theta}}(\mathbf{y}_1, ..., \mathbf{y}_T, S_1, ..., S_T)] | \mathbf{y}_1, ..., \mathbf{y}_T\} \\ =& E_{\boldsymbol{\theta}^{(l)}} \{\log[P_{\boldsymbol{\theta}}(S_1, ..., S_T)] | \mathbf{y}_1, ..., \mathbf{y}_T\} + \\ & E_{\boldsymbol{\theta}^{(l)}} \{\log[P_{\boldsymbol{\theta}}(\mathbf{y}_1, ..., \mathbf{y}_T | S_1, ..., S_T)] | \mathbf{y}_1, ..., \mathbf{y}_T\} \\ =& \left[\sum_{t=1}^{T} \sum_{i=1}^{M} \sum_{j \neq i} \frac{\xi_t(i,j)}{\gamma_t(i)} \log q_{ij} \right] + \left[\sum_{i=1}^{M} \gamma_0(i) \log \delta_i \right] + \\ & \left[\sum_{t=1}^{T} \sum_{j=1}^{M} \sum_{n=1}^{D} \frac{\eta_t(j,n)}{\gamma_t(i)} \log r_j(n) \right] + \\ & \left[\sum_{t=1}^{T} \sum_{j=1}^{M} \gamma_t(j) \log P(\mathbf{y}_t | \mathbf{y}_{t-1:\max(1,t-p)}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \mathbf{A}_j) \right], \end{aligned} \tag{8}$$

where $\boldsymbol{\theta}^{(l)}$ is the parameter value at the l^{th} iteration, and $P(\mathbf{y}_t | \mathbf{y}_{t-1:\max(1,t-p)}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \mathbf{A}_j)$ is the state-dependent density for the p^{th} order Gaussian autoregressive process.

In the M-step, we can harness the separability of parameters in $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)})$ to maximize each component individually as follows,

$$\delta_j = \gamma_0(j) / \sum_j \gamma_0(j), \tag{9}$$

$$q_{ij} = \sum_t \xi_t(i,j) / \sum_{j \neq i} \sum_t \xi_t(i,j), \tag{10}$$

$$r_j(n) = \sum_t \eta_t(j,n) / \sum_n \sum_t \eta_t(j,n). \tag{11}$$

Then $\boldsymbol{\mu}_j$ is updated as the unpenalized intercept in the weighted least squares regression for the VAR model with LASSO regularization, where each observation $\mathbf{y}_t|\mathbf{y}_{t-1:t-p}$ is weighted by $\gamma_t(j)$. The autoregression matrix $\mathbf{A}_j$ is updated as the coefficients in the same weighted least squares regression with LASSO regularization. These updates are carried out using coordinate descent algorithm detailed by Friedman et al. (2007).

$\boldsymbol{\Sigma}_j$ is updated as a convex combination of the weighted error variance from VAR and a scaled identity matrix with the same trace.

Asymptotic properties

The asymptotic properties for the maximum likelihood estimators in HMM under suitable regularity conditions have been proved successively in Leroux (1992), Bickel, Ritov, and Ryden (1998), Douc, Matias, and others (2001), Cappé, Moulines, and Rydén (2009), and An et al. (2013).

Furthermore, Barbu and Limnios (2009) (also in Trevezas and Limnios (2011)) extended proof for the consistency and asymptotic normality of the maximum likelihood estimators for finite-state discrete-time hidden semi-Markov models. The conditions and results are summarized as follows,

(B1) If for any states $i, j \in \{1, ..., M\}$, there is a positive integer τ such that $P(S_{t+\tau} = j|S_t = i) > 0$

(B2) The conditional state duration distributions $r_i(.)$ have finite support $\forall i \in \{1, ..., M\}$.

Under assumptions (B1) and (B2), the maximum likelihood estimator $\hat{\theta}_T$ is strongly consistent as $T \longrightarrow \infty$.

In the class of hidden semi-Markov model with a finite state space, assumption (B1) means that the Markov chain is irreducible. This holds when all the states communicate with each other, i.e. there is only one communication class in the transition matrix. (B2) automatically holds when we use the discrete nonparametric state duration distribution in the hidden semi-Markov model because we explicit assign probability mass to a finite collection of possible durations. In case a state duration density with infinite support is adopted, we can censor the distribution at a maximum duration D to satisfy the assumption.

Computational cost

To compute the likelihood in the E-step, Rabiner (1989) pointed out that the computational complexity $O(M^2T)$ for an M-state HMM with length T, and $O(M^2D^2T)$ for an M-state explicit duration HSMM censored at the largest duration D. Further in our VAR(p)-HSMM framework, the dimension of the observed series is d and the order of autoregression is p. Therefore, we have to include the computational cost of $O(d^3 + d^2p)$ to compute the multivariate normal density in each forward-backward variable. This adds to a total computational cost $O(M^2D^2T(d^3 + d^2p))$ in the E-step.

In the M-step, the most computationally expensive part is the update for the the autoregression matrices under the elastic net regularization. Based on the results from Friedman et al. (2007), the computational cost of the coordinate descent algorithm to solve LASSO is $O(Md^2pT)$ for M p^{th} order vector autoregressions of dimension d. This computational cost is dominated by that from the E-step.

Therefore, the total computational complexity is $O(M^2D^2T(d^3 + d^2p))$ for each EM iteration. As we can see, the algorithm scales linearly in the length of the series T and autoregression order p, but scales quadratically with the number of latent states M and the maximum censored duration D, and scales cubically with the dimension d.

Analysis on the NYSE portfolio data

We apply the proposed model on the New York Stock Exchange (NYSE) financial portfolio data, which consists of the daily closing price of 50 most active NYSE stocks from 2015-01-02 to 2016-12-30 so that each time series is of length 504. This data set is publicly available for download in the R package "rarhsmm". We use the log return as the observed multivariate sequence $\{\mathbf{y}_t\}$ with dimension 50 such that

$$y_t = \log \frac{\text{price}_{t+1}}{\text{price}_t} \quad t = 1, ..., 503,$$

Our analysis shows there is a fairly strong, positive correlation in the lag 0 log returns among most of the 50 stocks. In contrast, the right panel displays the lag 1 correlation matrix, which is rather sparse. Indeed, 83 of the lag 1 sample correlations are significantly different from zero after testing by Fisher z-transformation ($p < 0.05$). This sparsity motivates the use of regularized estimators on the state-dependent autoregression matrices in the VAR(p)-HSMM framework.

The model selection is performed among the competing regularized models [VAR, HMM, VAR(p)-HSMM] using the minimum MSFE criterion. The first 303 observations were used for training, the next 100 for validation, and the final 100 for forecasting. We set 15 grid points that fall with equal space on the log scale between 0.0001 and 1 for LASSO parameter on VAR coefficients. Similarly, we set 15 grid points that fall with equal space on the log scale between 0.1 and 100 for the shrinkage on the covariances. When fitting the VAR-HSMMs, the maximum latent state duration is set to be 30 days and all latent state duration densities are chosen to be discrete nonparametric. From Table 1, all competing models perform comparably well in terms of the MSFE. Both regularized VAR(1)-HSMM and VAR(2)-HSMM with 2 states achieved the lowest MSFE of 2.271. Thus, the regularized VAR(1)-HSMM is selected to be the final model since it is more parsimonious.

The scatter plot in Figure 1 depicts the log returns of the 50 stocks from 2015-01-02 to 2016-12-30. A sequence of the decoded latent states using Viterbi algorithm is overlaid on top of the scatter plot. We can see that state 2 corresponds to the period with a higher volatility in the log return of the 50 stocks while state 1 represents a relatively stable economic period. Figure 2 and Figure 3 display the scatter plot and empirical distributions for the fitted means and variances in the two latent states (stable versus volatile). In Figure 2, we can see that the means in both states are centered around 0, but the spread in the means of state 2 is much larger than that in state 1. In Figure 3, it seems that most of the stocks have a larger variance for log return in state 2 than in state 1

Model ID	Model specification	MSFE
1	Regularized VAR(1)	2.293
2	Regularized HMM with 2 latent states	2.288
3	Regularized VAR(1)-HSMM with 2 latent states	2.271
4	Regularized VAR(2)-HSMM with 2 latent states	2.271
5	Regularized VAR(1)-HSMM with 3 latent states	2.289

Table 1: Summary of model selection on the NYSE portfolio data. The regularization parameters are selected using cross-validation by minimizing one-step-ahead mean-square forecast error (MSFE). We set 15 grid points that fall with equal space on the log scale between 0.0001 and 1 for LASSO parameter on VAR coefficients. Similarly, we set 15 grid points that fall with equal space on the log scale between 0.1 and 100 for the shrinkage on the covariances.

since the majority of the points lie above the 45 degree line. This result also corroborates the claim that state 2 stands for a more volatile economy than state 1.

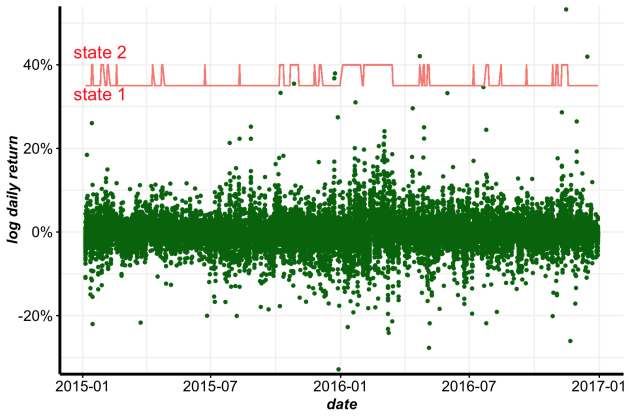

Figure 1: Log returns and the decoded latent states for the 50 stocks in the NYSE portfolio from 2015-01-02 to 2016-12-30.

Discussions

The class of regularized VAR-HSMM provides a flexible framework to model the switching data generating regimes in multivariate financial time series data, which can work especially well when these state-dependent covariance and autoregression matrices are indeed sparse. In the case study in Section 4, the maximum latent duration (D) is set to be 30 days so as to account for the potential long temporal dependence. We do not want D to be too small, in which case the VAR(p)-HSMM would boil down to VAR(p)-HMM. Although the computation cost of the algorithm increases quadratically in D, the number of parameters only increases linearly in D. In the final regularized model of VAR(1)-HSMM, there are 2909 estimated parameters that are nonzero, where 2550 of them belong to the state-dependent covariance matrices. The fitted means in both states are centered around zero, and there exists strong, positive correlation among most of the stocks in both states. However, the financial returns in state 1 (stable) seems to satisfy the white noise assumption while there is some evidence of lag 1 correlation in state 2 (volatile).

In addition, there are other choices of regularization on the covariance and autoregression matrices. For instance, graphical LASSO (Yuan and Lin, 2007) could be used on the state-dependent covariance matrices and SCAD (Fan and Li, 2001) could be used on the autoregression matrices, which is the strategy adopted by Monbet and Ailliot (2017) in their VAR-HMM. Another common technique to reduce the number of parameters in covariance and autoregression matrices is to make parametric assumptions on their structures, which will in turn require testing the goodness-of-fit for those assumptions.

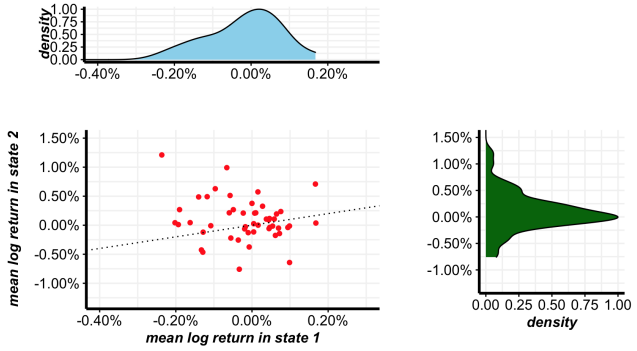

Figure 2: Scatter plot and empirical distributions of the fitted means for the log returns in state 1 (stable) and 2 (volatile).

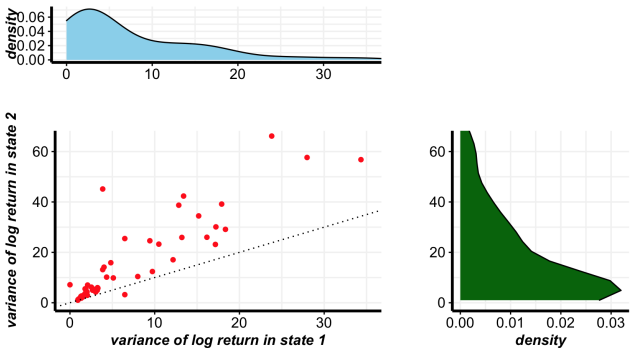

Figure 3: Scatter plot and empirical distributions of the fitted variances for the log returns in state 1 (stable) and 2 (volatile).

References

An, Y.; Hu, Y.; Hopkins, J.; and Shum, M. 2013. Identifiability and inference of hidden markov models. Technical report, Citeseer.

Andersen, T. G.; Bollerslev, T.; Diebold, F. X.; and Ebens, H. 2001. The distribution of realized stock return volatility. *Journal of financial economics* 61(1):43–76.

Bańbura, M.; Giannone, D.; and Reichlin, L. 2010. Large bayesian vector auto regressions. *Journal of Applied Econometrics* 25(1):71–92.

Barbu, V. S., and Limnios, N. 2009. *Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis*, volume 191. Springer Science & Business Media.

Baum, L. E.; Petrie, T.; Soules, G.; and Weiss, N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics* 41(1):164–171.

Bickel, P. J.; Ritov, Y.; and Ryden, T. 1998. Asymptotic normality of the maximum-likelihood estimator for general hidden markov models. *Annals of Statistics* 1614–1635.

Bulla, J., and Bulla, I. 2006. Stylized facts of financial time series and hidden semi-markov models. *Computational Statistics & Data Analysis* 51(4):2192–2209.

Cappé, O.; Moulines, E.; and Rydén, T. 2009. Inference in hidden markov models. In *Proceedings of EUSFLAT Conference*, 14–16.

De Pierro, A. R. 1995. A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE transactions on medical imaging* 14(1):132–137.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* 1–38.

Ding, Z.; Granger, C. W.; and Engle, R. F. 1993. A long memory property of stock market returns and a new model. *Journal of empirical finance* 1(1):83–106.

Douc, R.; Matias, C.; et al. 2001. Asymptotics of the maximum likelihood estimator for general hidden markov models. *Bernoulli* 7(3):381–420.

Fan, J., and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* 96(456):1348–1360.

Fan, J., and Yao, Q. 2017. *The Elements of Financial Econometrics*. Cambridge University Press.

Fiecas, M.; Franke, J.; von Sachs, R.; and Tadjuidje Kamgaing, J. 2017. Shrinkage estimation for multivariate hidden markov models. *Journal of the American Statistical Association* 112(517):424–435.

Francq, C., and Zakoıan, J.-M. 2001. Stationarity of multivariate markov–switching arma models. *Journal of Econometrics* 102(2):339–364.

Franke, J.; Härdle, W. K.; and Hafner, C. M. 2004. *Statistics of financial markets*, volume 2. Springer.

Friedman, J.; Hastie, T.; Höfling, H.; Tibshirani, R.; et al. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics* 1(2):302–332.

Granger, C. W. J., and Ding, Z. 1995. Some properties of absolute return: An alternative measure of risk. *Annales d'Economie et de Statistique* 67–91.

Green, P. J. 1990. On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* 443–452.

Hamilton, J. D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society* 357–384.

Ledoit, O., and Wolf, M. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis* 88(2):365–411.

Leroux, B. G. 1992. Maximum-likelihood estimation for hidden markov models. *Stochastic processes and their applications* 40(1):127–143.

Monbet, V., and Ailliot, P. 2017. Sparse vector markov switching autoregressive models. application to multivariate time series of temperature. *Computational Statistics & Data Analysis* 108:40–51.

Nicholson, W. B.; Matteson, D. S.; and Bien, J. 2014. Structured regularization for large vector autoregression. *Cornell University*.

Rabiner, L. R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Rydén, T.; Teräsvirta, T.; and Åsbrink, S. 1998. Stylized facts of daily return series and the hidden markov model. *Journal of applied econometrics* 217–244.

Sancetta, A. 2008. Sample covariance shrinkage for high dimensional dependent data. *Journal of Multivariate Analysis* 99(5):949–967.

Städler, N., and Mukherjee, S. 2013. Penalized estimation in high-dimensional hidden markov models with state-specific graphical models. *The Annals of Applied Statistics* 2157–2179.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Trevezas, S., and Limnios, N. 2011. Exact mle and asymptotic properties for nonparametric semi-markov models. *Journal of Nonparametric Statistics* 23(3):719–739.

Yang, M. 2000. Some properties of vector autoregressive processes with markov-switching coefficients. *Econometric Theory* 16(1):23–43.

Yu, S.-Z. 2010. Hidden semi-markov models. *Artificial intelligence* 174(2):215–243.

Yuan, M., and Lin, Y. 2007. Model selection and estimation in the gaussian graphical model. *Biometrika* 94(1):19–35.