A Regularized Vector Autoregressive Hidden Semi-Markov Model

with application to Multivariate Financial Data

Zekun Xu & Ye Liu

Statistics Department, North Carolina State University

Abstract

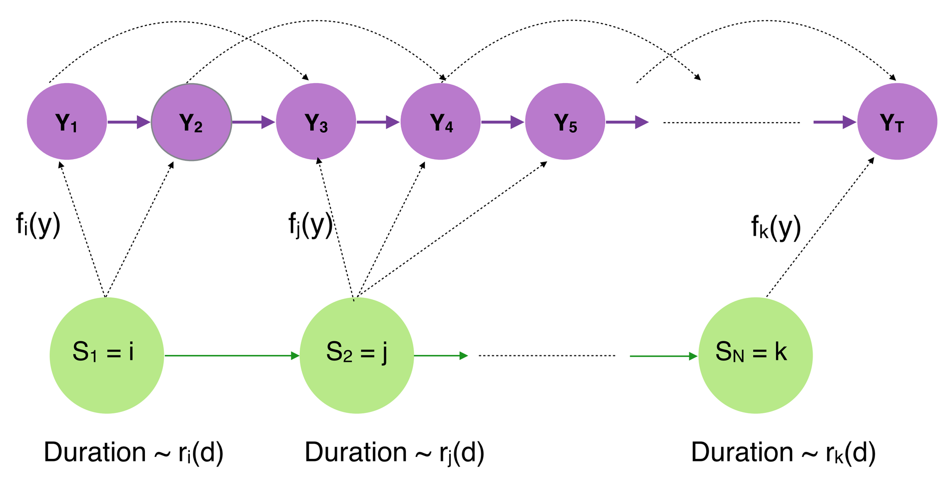
We provide a flexible p^{th} order vector autoregressive hidden semi-Markov model (VAR(p)-HSMM) framework to analyze multivariate financial time series with switching data generating regimes. Furthermore, we enhance the EM algorithm to stabilize the parameter estimation by embedding regularized estimators for the state-dependent covariance matrices and autoregression matrices in the M-step. Simulation studies are carried out to evaluate the performance of our proposed regularized estimators. In addition, we demonstrate the use of a regularized VAR(p)-HSMM to model the real NYSE financial portfolio data.

Introduction

In finance and economics, time series often have more than one latent data generating mechanisms. For example, it is reasonable to assume the performance of a financial portfolio during a bull market to follow a very different autoregressive process from that during a bear market. As a result, the class of hidden Markov models (HMM) arise as a natural solution to analyze time series with switching data generating regimes. HMM is a bivariate discrete time stochastic process $\{S_t, Y_t\}_{t \geq 0}$ such that

1. $\{S_t\}$ is a Markov chain, i.e. $P(S_t|S_{t-1}, ..., S_1) = P(S_t|S_{t-1})$
2. $\{Y_t\}$ are conditionally independent given $\{S_t\}$

In practice, the 2 assumptions are both too strong to hold for financial time series. To generalize assumption 1, the class of hidden semi-Markov models (HSMM) allows for explicitly modelling the time duration of the hidden states rather than assume a memoryless geometric distribution. In the meantime, assumption 2 can be dropped in the class of Markov-switching models, which incorporates state-dependent Gaussian autoregressive processes, also known as autoregressive hidden Markov models (ARHMM). 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.



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 n has $\frac{Mn(n+1)}{2}$ parameters in the state-dependent covariance matrices and Mpn^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 n is moderate. Therefore, regularizations are needed to stabilize the parameter estimation.

In this project, we provide a detailed parameter estimation procedure for a regularized VAR(p)-HSMM, where we integrated the elastic net regularization on the autoregression matrices and shrinkage regularization on the covariance matrices into the EM algorithm for parameter estimation. Our R package **"rarhsmm"** has been developed for fitting regularized VAR(p)-HSMM, which is available at `https://cran.r-project.org/web/packages/rarhsmm/index.html`

Methodology

Modelling framework for VAR(p)-HSMM

- Let M be the number of latent states.
- An initial state, $S_1 = i(i \in 1, ..., M)$ is chosen according to the initial state distribution δ_i.
- A duration d_1 is chosen according to the nonparametric state duration density $r_i(d_1)$, which is censored at a maximum duration D.
- Observations $\mathbf{y}_1, ..., \mathbf{y}_{d1} \in \mathbb{R}^n$ are chosen according to the state-dependent p^{th} order Gaussian vector autoregressive process

$$\mathbf{y}_t = \mu(S_t) + \sum_{k=1}^{p} \mathbf{A}_k(S_t)\mathbf{y}_{t-k} + \mathbf{\Sigma}(S_t) \quad t = 1, ..., d_1$$

  where $\boldsymbol{\mu}(S_t)$, $\mathbf{\Sigma}(S_t)$, and $\mathbf{A}_k(S_t)$ are the conditional mean, covariance matrix, and k^{th}-order autoregression matrices conditioning on S_t.
- The next state, $S_2 = j$, is chosen according to the state transition probabilities, q_{ij}

Parameter estimation: a modified EM algorithm

In the E-step of the l^{th} iteration, we define and compute

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}) = E_{\boldsymbol{\theta}^{(l)}}\{\log[P_{\boldsymbol{\theta}}(Y_1, ..., Y_T, S_1, ..., S_T)]|y_1, ..., y_n\}$$

In the M-step, except for the covariance matrices $\mathbf{\Sigma}_j$ and autoregression matrices $\mathbf{A}_j$ for $j = 1, ..., M$, we update all the other parameters by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)})$.

The regularized estimator for state-dependent covariance matrices is a convex combination of the maximum likelihood estimator and a scaled identity matrix with the same trace

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

where $\lambda_\Sigma \geq 0$ controls the strength of the regularization. Note that when $\lambda_\Sigma = 0$, we have $\mathbf{\Sigma}^r = \hat{\mathbf{\Sigma}}^{mle}$. This regularized estimator yields an invertible and well-conditioned covariance matrix even when the sample covariance matrix is close to singularity. This regularization has a Bayesian analogy where $\mathbf{\Sigma}^r$ can be considered as the combination of prior information (centered around $c\mathbf{I}$) and sample information (centered around $\hat{\mathbf{\Sigma}}^{mle}$).

The regularized estimator for state-dependent autoregressive coefficients is based on the elastic net regularization such that

$$\mathbf{a}^r = \arg\min_{\mathbf{a}} \| \operatorname{vec}(Y_{p+1:T}) - \mu + \sum_{k=1}^{p} \mathbf{a}_k^{\intercal} \operatorname{vec}(Y_{p+1-k:T-k})\|_2^2 + \lambda_a[\alpha\|\mathbf{a}\|_1 + (1-\alpha)\|\mathbf{a}\|_2^2]$$

where $\mathbf{a} = [\mathbf{a}_p^{\intercal}, ..., \mathbf{a}_1^{\intercal}]^{\intercal} = [\operatorname{vec}(\mathbf{A}_p)^{\intercal}, ..., \operatorname{vec}(\mathbf{A}_1)^{\intercal}]^{\intercal}$ is the vectorization of the state-dependent autoregression matrices. Here $\lambda_a \geq 0$ controls the strength of the regularization, while α adjusts for the mixing weight of ℓ_1 and ℓ_2 penalty. The elastic net regularization is an improvement on LASSO in that it enables strongly correlated predictors to stay in or drop out of the model together. A coordinate descent algorithm is used to solve the convex optimization problem of elastic net shrinkage.

Simulation Results

In order to evaluate the performance of our regularized estimator on the state-dependent autoregressive coefficients and covariance matrices, we simulated VAR(1)-HSMM series of length 500 with 2 latent states under the following scenarios:

1. Dimension = 10; sparse covariance and autoregression matrices
2. Dimension = 10; dense covariance and autoregression matrices
3. Dimension = 50; sparse covariance and autoregression matrices
4. Dimension = 50; dense covariance and autoregression matrices

The two competing models are as follows:

1. Model 1 (not regularized): $\lambda_a = \lambda_\Sigma = 0$
2. Model 2 (regularized): $\lambda_a = \lambda_\Sigma = 1, \quad \alpha = 0.8$

Parameter	**Sparse Matrices**		**Dense Matrices**	
	Model 1	Model 2	Model 1	Model 2
Dimension = 10				
$\|\mathbf{A}_1^\star - \hat{\mathbf{A}}_1\|_F$	0.78	0.28	0.80	1.17
$\|\mathbf{A}_2^\star - \hat{\mathbf{A}}_2\|_F$	0.99	0.24	0.97	1.11
$\|\mathbf{\Sigma}_1^\star - \hat{\mathbf{\Sigma}}_1\|_F$	0.82	0.57	0.85	1.30
$\|\mathbf{\Sigma}_2^\star - \hat{\mathbf{\Sigma}}_2\|_F$	0.83	0.47	0.85	0.69
Dimension = 50				
$\|\mathbf{A}_1^\star - \hat{\mathbf{A}}_1\|_F$	3.35	1.27	3.56	3.61
$\|\mathbf{A}_2^\star - \hat{\mathbf{A}}_2\|_F$	7.01	1.30	4.42	3.95
$\|\mathbf{\Sigma}_1^\star - \hat{\mathbf{\Sigma}}_1\|_F$	3.05	1.97	5.13	4.62
$\|\mathbf{\Sigma}_2^\star - \hat{\mathbf{\Sigma}}_2\|_F$	3.05	2.12	5.13	5.38

Table 1: Mean difference in Frobenius norm between the true values and estimates via 1000 simulations. Standard Error is in the range of (0.01,0.1). Model 2 is uniformly better than model 1 in the case of sparse covariance and autoregression matrices, which is an ideal situation for regularized estimators.

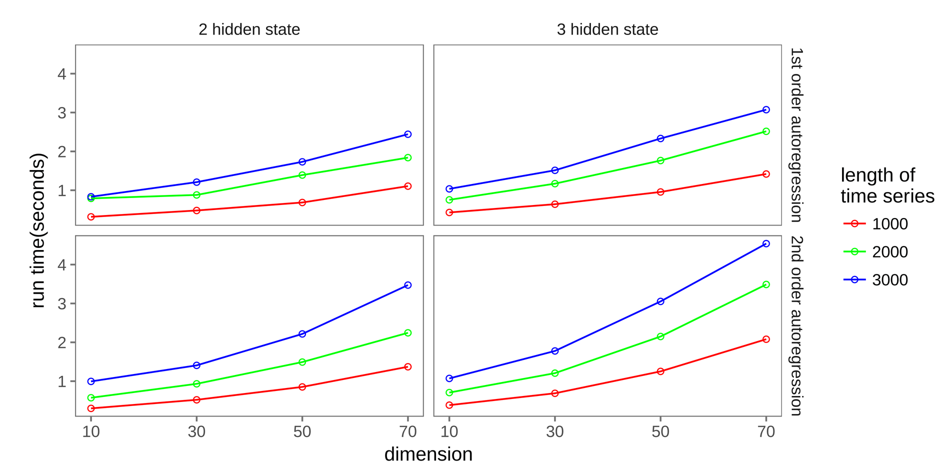


Figure 1: average running time of a single EM update in the regularized estimation algorithm in different problem sizes on a 2.7 GHz Intel Core i5 processor.

Empirical Result

The financial portfolio data consists of the log daily return of 50 NYSE stocks from 2015-01-02 to 2016-12-30 so that each time series is of length 503. Using the minimum AIC criterion for model selection, our final model is a 2-state VAR(1)-HSMM.

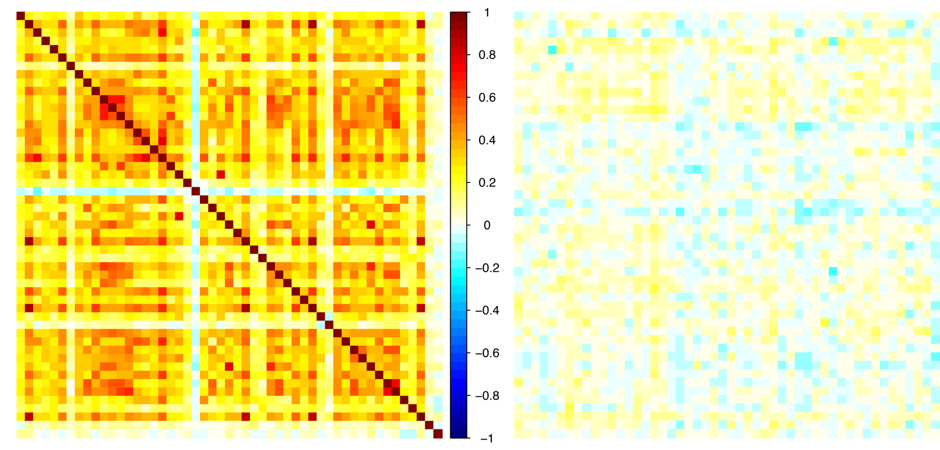


Figure 2: The left panel shows there is a fairly strong positive correlation in most of the lag 0 log returns among the 50 stocks. In contrast, the right panel displays the lag 1 autocorrelation matrix, which is rather sparse. A sparse autocorrelation justifies the use of regularized estimators.

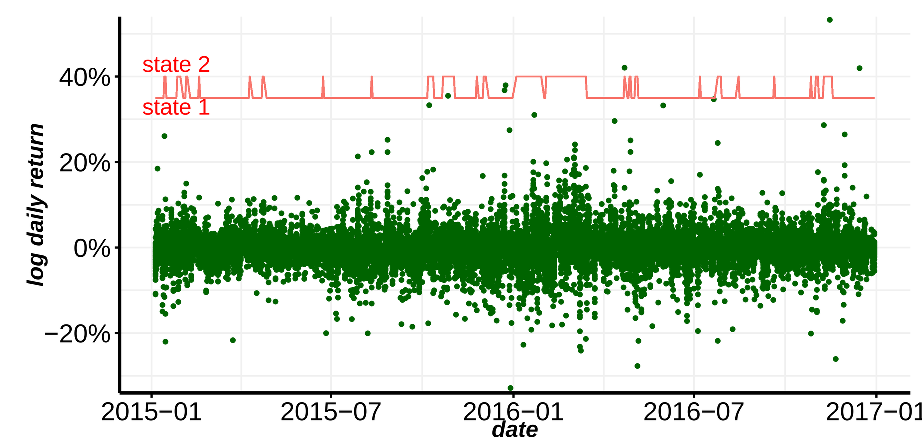


Figure 3: The scatter plot depicts the log returns of the 50 stocks from 2015-01-02 to 2016-12-30. A sequence of the 2 decoded latent states is overlaid on top of the scatter plot. We can see that state 2 corresponds to the period with a higher volatility.

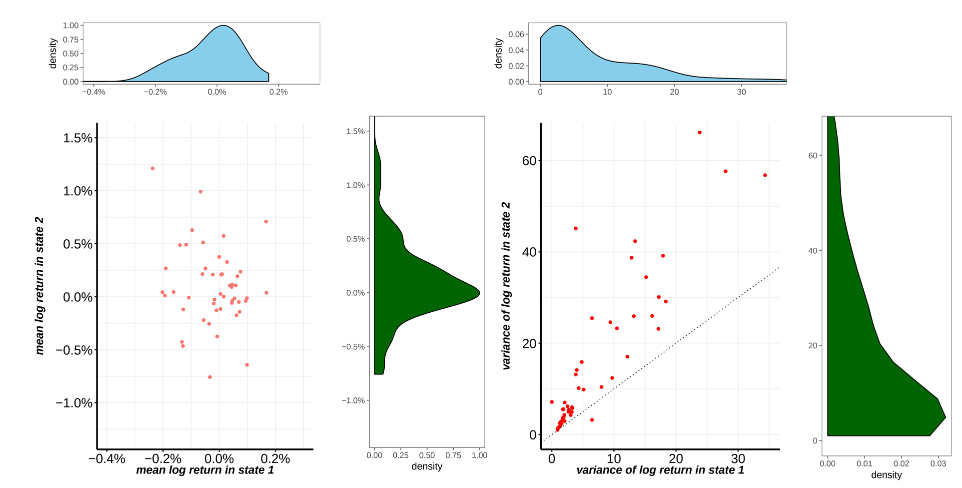


Figure 4: The left panel shows the mean log returns of the 50 stocks in state 2 versus those in state 1. Although the means in both states are centered around 0, the spread in means of state 2 is much larger than that in state 1. The right panel displays the variances in the log returns of the 50 stocks in state 2 versus state 1. Since the majority of the points lie above the 45 degree line, it seems that the variance in state 2 is greater than that in state one for most of the stocks.

Conclusions

- VAR(p)-HSMM provides a flexible framework to model the switching data generating regimes in multivariate financial time series data.
- A regularized VAR(p)-HSMM can yield stable estimates for the state-dependent covariance and autoregression matrices. The regularized estimators work especially well when these matrices are indeed sparse.