GPU-Accelerated Bayesian Learning in Simultaneous Graphical Dynamic Linear Models

Lutz Gruber[1,2], Mike West[1]

Second Bayesian Young Statisticians Meeting (BAYSM 2014)
Vienna, September 18–19, 2014

[1] Duke University, Department of Statistical Sciences, Durham, NC, USA
<lg145,mw@stat.duke.edu>

[2] Technische Universität München, München, Germany

Abstract

We introduce a novel class of Simultaneous Graphical Dynamic Linear Models (SGDLMs) for learning and prediction of increasingly high-dimensional time series. The multivariate time series are decoupled into a parallel set of univariate dynamic linear models using a Variational Bayes strategy. At each time point, parallel sequential updating and forecasting for each series applies. Importance sampling recovers the exact multivariate posterior. Model evolution to the next time point uses a Bayes' strategy to decouple and proceed. Our GPU implementation, using C++/CUDA, exploits massive parallelization for decoupled analyses and simulation. The overall modeling and computational strategy is enormously scalable with time series dimension. We demonstrate that this allows for an involved, real-time Bayesian analysis of a 400-dimensional daily stock return time series in portfolio investment studies.

Keywords: high-dimensional data, time series; forecasting; variational bayes; importance sampling

1 Introduction

The dynamic linear model (DLM) is an established Bayesian time series model [10, 7]. The model's versatility makes it a popular choice in many fields such as finance, econometrics, biology and genomics.

We present the simultaneous graphical dynamic linear model (SGDLM) as a sparser multivariate extension of the univariate DLM than the standard multivariate DLM, which uses a variance-discounting Wishart prior. The number of parameters of the standard DLM increase quadratically with the number of series. We hope to make forecasting of high-dimensional time series more robust by reducing the number of parameters dramatically. Previous attempts to sparser versions of the multivariate DLM include [11, 1, 3, 8]. Furthermore, the SGDLM allows forward filtering in parallel, thus enabling efficient massively parallel implementations for GPUs.

We provide a GPU implementation in C++/CUDA along with the presentation of the new model. GPU computation is ideally suited for parallel analysis, and the overall modeling and computational strategy is enormously scalable with time series dimension as a result [6, 9, 4].

2 Model Specification

Let $\mathbf{y}_t = (y_{1t}, \ldots, y_{mt})'$, $t = 1:T$ be an m-variate time series whose time t observation $\mathbf{y}_t$ follows an m-variate normal distribution $N(\mathbf{A}_t\boldsymbol{\mu}_t, \boldsymbol{\Omega}_t^{-1})$. The series $j = 1 : m$ are modeled as simultaneously coupled univariate dynamic linear models (DLMs; [10, 7]):

$$\begin{aligned} y_{jt} &= \mathbf{x}'_{jt}\phi_{jt} + \gamma'_{jt}\mathbf{y}_{sp(j),t} + \nu_{jt} \\ &= \mathbf{F}'_{jt}\boldsymbol{\theta}_{jt} + \nu_{jt}. \end{aligned}$$

We write $sp(j) \subseteq \{1 : m\} \setminus \{j\}$ for the j-th parental set, which denotes the simultaneous parents of the j-th series; γ_{jt} is the vector of dynamic regression coefficients γ_{jht}, $h \in sp(j)$, with dimension $p_{j\gamma} = |sp(j)|$; $\mathbf{x}_{jt} \in \mathbb{R}^{p_{j\phi}}$ is a known column vector of predictors or constants, with corresponding dynamic regression coefficients in the column state vector $\phi_{jt} \in \mathbb{R}^{p_{j\phi}}$. The observation noise terms $\nu_{jt} \sim N(0, \lambda_{jt}^{-1})$ are independent across series and over time; $\boldsymbol{\nu}_t = (\nu_{1t}, \ldots, \nu_{mt})' \sim N(\mathbf{0}, \boldsymbol{\Lambda}_t^{-1})$ with precision matrix $\boldsymbol{\Lambda}_t = \text{diag}(\lambda_{1t}, \ldots, \lambda_{mt})$. We set $\gamma_{jht} = 0$ for each $h \notin sp(j)$ and collect the effective coefficients γ_{jt} along with the implicit zero values in the matrix

$$\boldsymbol{\Gamma}_t = \begin{pmatrix} 0 & \gamma_{1,2,t} & \gamma_{1,3,t} & \cdots & \gamma_{1,m,t} \\ \gamma_{2,1,t} & 0 & \gamma_{2,3,t} & \cdots & \gamma_{2,m,t} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{m-1,1,t} & \cdots & \gamma_{m-1,m-2,t} & 0 & \gamma_{m-1,m,t} \\ \gamma_{m,1,t} & \gamma_{m,2,t} & \cdots & \gamma_{m,m-1,t} & 0 \end{pmatrix}.$$

Sparsity can be achieved in that $sp(j)$ may be a few, or no, elements. A sparse model has a sparse coefficient matrix $\boldsymbol{\Gamma}_t$.

Writing $\boldsymbol{\mu}_t = (\mu_{1t}, \ldots, \mu_{mt})'$ with $\mu_{jt} = \mathbf{x}'_{jt}\phi_{jt}$, $\mathbf{A}_t = (\mathbf{I} - \boldsymbol{\Gamma}_t)^{-1} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{\Omega}_t = (\mathbf{I} - \boldsymbol{\Gamma}'_t)\boldsymbol{\Lambda}_t(\mathbf{I} - \boldsymbol{\Gamma}_t) \in \mathbb{R}^{m\times m}$, we obtain

$$(\mathbf{I} - \boldsymbol{\Gamma}_t)\mathbf{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\nu}_t, \tag{1}$$

$$\mathbf{y}_t \sim N(\mathbf{A}_t\boldsymbol{\mu}_t, \boldsymbol{\Omega}_t^{-1}). \tag{2}$$

The across-series dependence characteristics of the covariance matrix $\boldsymbol{\Sigma}_t = \boldsymbol{\Omega}^{-1}$ arise from the pattern of the simultaneous parents in $\boldsymbol{\Gamma}_t$. When the entries of $\boldsymbol{\Gamma}_t$ are reduced to being either zero or non-zero, it can be interpreted as the adjacency matrix of an equivalent graphical model, hence the term "simultaneous graphical dynamic linear model" (SGDLM).

3 Forward Filtering

Initial Prior We choose independent normal-inverse gamma priors to reflect the initial information $\mathcal{D}_0$; these are conjugate for univariate DLMs [10]:

$$p(\boldsymbol{\Theta}_1, \boldsymbol{\Lambda}_1|\mathcal{D}_0) = \prod_{j=1:m} N(\boldsymbol{\theta}_{j1}|\mathbf{a}_{j1}, \mathbf{R}_{j1}/(c_{j1}\lambda_{j1}))\ G(\lambda_{j1}|r_{j1}/2, r_{j1}c_{j1}/2), \tag{3}$$

with prior parameters $\mathbf{a}_{j1} \in \mathbb{R}^{p_j}$, $\mathbf{R}_{j1} \in \mathbb{R}^{p_j\times p_j}$, $r_{j1} > 0$ and $c_{j1} > 0$.

Posterior at Time t The multivariate joint posterior of $\mathbf{\Theta}_t$ and $\mathbf{\Lambda}_t$ given information $\mathcal{D}_t$ has density

$$p(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t) \propto |\mathbf{I} - \mathbf{\Gamma}_t| \prod_{j=1:m} N(\boldsymbol{\theta}_{jt}|\mathbf{m}_{jt}, \mathbf{C}_{jt}/(s_{jt}\lambda_{jt}))\ G(\lambda_{jt}|n_{jt}/2, n_{jt}s_{jt}/2). \qquad (4)$$

The parameters $\mathbf{m}_{jt} \in \mathbb{R}^{p_j}$, $\mathbf{C}_{jt} \in \mathbb{R}^{p_j \times p_j}$, $n_{jt} > 0$ and $s_{jt} > 0$ are obtained by executing the posterior step independently within each series $j = 1 : m$ [10]:

- $Q_{jt} = c_{jt} + \mathbf{F}'_{jt}\mathbf{R}_{jt}\mathbf{F}_{jt}$;
- $\mathbf{A}_{jt} = \mathbf{R}_{jt}\mathbf{F}_{jt}/Q_{jt}$;
- $e_{jt} = y_{jt} - \mathbf{F}'_{jt}\mathbf{a}_{jt}$;
- $n_{jt} = r_{jt} + 1$;
- $s_{jt} = c_{jt}(r_{jt} + e^2_{jt}/Q_{jt})/(r_{jt} + 1)$;
- $\mathbf{m}_{jt} = \mathbf{a}_{jt} + \mathbf{A}_{jt}e_{jt}$;
- $\mathbf{C}_{jt} = (\mathbf{R}_{jt} - \mathbf{A}_{jt}\mathbf{A}'_{jt}Q_{jt})s_{jt}/c_{jt}$.

The marginal posteriors are tied together, or "recoupled", only by the determinant of $(\mathbf{I} - \mathbf{\Gamma}_t)$ as corrective factor of the posterior density (Eq. 4). Observing this detail is crucial to developing a mostly parallel computational implementation of the posterior estimation step.

Step-Ahead Prior at Time t To obtain an analytically tractable result, we assume that the posterior parameters $(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t)$ are decoupled to follow an independent normal inverse gamma distribution with parameters $\widetilde{\mathbf{m}}_{jt}$, $\widetilde{\mathbf{C}}_{jt}$, $\widetilde{n}_{jt}$ and $\widetilde{s}_{jt}$:

$$p(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t) \approx \prod_{j=1:m} N(\boldsymbol{\theta}_{jt}|\widetilde{\mathbf{m}}_{jt}, \widetilde{\mathbf{C}}_{jt}/(\widetilde{s}_{jt}\lambda_{jt}))\ G(\lambda_{jt}|\widetilde{n}_{jt}/2, \widetilde{n}_{jt}\widetilde{s}_{jt}/2). \qquad (5)$$

Then the one step ahead prior distribution $(\mathbf{\Theta}_{t+1}, \mathbf{\Lambda}_{t+1}|\mathcal{D}_t)$ follows as a normal-inverse gamma distribution with parameters $\mathbf{a}_{j,t+1} \in \mathbb{R}^{p_j}$, $\mathbf{R}_{j,t+1} \in \mathbb{R}^{p_j \times p_j}$, $r_{j,t+1}$ and $c_{j,t+1}$ that is independent across the series $j = 1 : m$:

$$\begin{aligned} &p(\mathbf{\Theta}_{t+1}, \mathbf{\Lambda}_{t+1}|\mathcal{D}_t) \\ &= \prod_{j=1:m} N(\boldsymbol{\theta}_{j,t+1}|\mathbf{a}_{j,t+1}, \mathbf{R}_{j,t+1}/(c_{j,t+1}\lambda_{j,t+1}))G(\lambda_{j,t+1}|r_{j,t+1}/2, r_{j,t+1}c_{j,t+1}/2). \end{aligned} \qquad (6)$$

The parameters are

- $r_{j,t+1} = \beta_j\widetilde{n}_{jt}$;
- $c_{j,t+1} = \widetilde{s}_{jt}$;
- $\mathbf{a}_{j,t+1} = \mathbf{G}_{j,t+1}\widetilde{\mathbf{m}}_{jt}$;
- $\mathbf{R}_{j,t+1} = \mathbf{G}_{j,t+1}\mathbf{\Delta}_j\widetilde{\mathbf{C}}_{jt}\mathbf{\Delta}'_j\mathbf{G}'_{j,t+1}$; this implicitly defines $\mathbf{W}_{j,t+1}$.

4 Variational Bayes Posterior Decoupling and Forecasting

The time evolution step (Section 3) requires the posterior distribution of the parameters $\mathbf{\Theta}_t$ and $\mathbf{\Lambda}_t$ be independent across the series $j = 1 : m$. Since this is not the case for the exact posterior distribution (Eq. 4), we adopt a Variational Bayes approximation.

4.1 Variational Bayes Decoupling of the Posterior

A Variational Bayes approach estimates the parameters of a pre-specified posterior distribution class to best approximate the true, intractably complicated exact posterior distribution [2].

The exact posterior distribution of $\mathbf{\Theta}_t$ and $\mathbf{\Lambda}_t$ is given in Equation 4. We expect the value of the determinant $|\mathbf{I}-\mathbf{\Gamma}_t|$ not to spread far from unity, given the sparse matrix $\mathbf{\Gamma}_t$. As a result, the exact posterior distribution will be close to an across series-independent normal inverse gamma distribution,

$$p(\mathbf{\Theta}_t, \mathbf{\Lambda}_t|\mathcal{D}_t) \approx p(\widetilde{\mathbf{\Theta}}_t, \widetilde{\mathbf{\Lambda}}_t) = \prod_{j=1:m} N(\widetilde{\boldsymbol{\theta}}_{jt}|\widetilde{\mathbf{m}}_{jt}, \widetilde{\mathbf{C}}_{jt}/(\widetilde{s}_{jt}\widetilde{\lambda}_{jt}))\ G(\widetilde{\lambda}_{jt}|\widetilde{n}_{jt}/2, \widetilde{n}_{jt}\widetilde{s}_{jt}/2). \quad (7)$$

The parameters $\widetilde{\mathbf{m}}_{jt} \in \mathbb{R}^{p_j}$, $\widetilde{\mathbf{C}}_{jt} \in \mathbb{R}^{p_j \times p_j}$, $\widetilde{n}_{jt} > 0$ and $\widetilde{s}_{jt} > 0$ are set to minimize the Kullback-Leibler divergence from the decoupled Variational Bayes posterior to the true multivariate posterior:

- $\widetilde{\mathbf{m}}_{jt} = \dfrac{E(\lambda_{jt}\boldsymbol{\theta}_{jt})}{E(\lambda_{jt})}$;
- $\widetilde{\mathbf{V}}_{jt} = E(\lambda_{jt}(\boldsymbol{\theta}_{jt} - \widetilde{\mathbf{m}}_{jt})(\boldsymbol{\theta}_{jt} - \widetilde{\mathbf{m}}_{jt})')$;
- $\widetilde{Q}_{jt} = E(\lambda_{jt}(\boldsymbol{\theta}_{jt} - \widetilde{\mathbf{m}}_{jt})'\widetilde{\mathbf{V}}_{jt}^{-1}(\boldsymbol{\theta}_{jt} - \widetilde{\mathbf{m}}_{jt}))$;
- $\widetilde{n}_{jt}$ such that $\log(\widetilde{n}_{jt} + p_j - \widetilde{Q}_{jt}) - \psi(\widetilde{n}_{jt}/2) - (p_j - \widetilde{Q}_{jt})/\widetilde{n}_{jt} - \log(2E(\lambda_{jt})) + E(\log \lambda_{jt}) = 0$;
- $\widetilde{s}_{jt} = \dfrac{\widetilde{n}_{jt} + p_j - \widetilde{Q}_{jt}}{\widetilde{n}_{jt}E(\lambda_{jt})}$;
- $\widetilde{\mathbf{C}}_{jt} = \widetilde{s}_{jt}\widetilde{\mathbf{V}}_{jt}$.

The parameters depend on some moments of the exact posterior. We will evaluate these expectations using importance sampling [5].

The overall Kullback-Leibler divergence of the approximation can be shown to be an upper bound for the Kullback-Leibler divergence of any series. Each series' approximation will be sufficiently good as long as the quality of the overall approximation is sufficient.

4.2 One Step Ahead Forecast at Time t

The observation equation (Eq. 1) can be written as

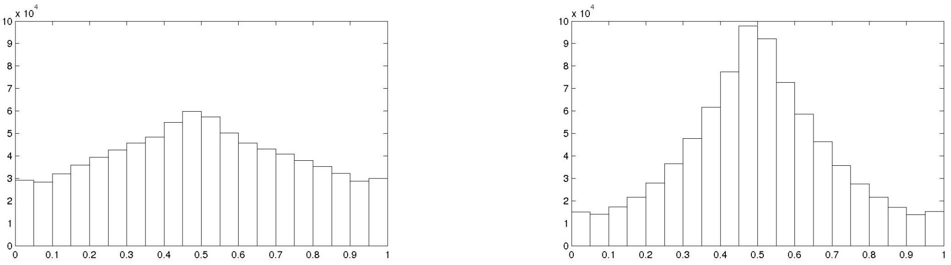
$$\mathbf{y}_{t+1} = (\mathbf{I} - \mathbf{\Gamma}_{t+1})^{-1}(\boldsymbol{\mu}_{t+1} + \boldsymbol{\nu}_{t+1}). \quad (8)$$

There is no closed-form analytical expression of the forecast distribution available, given the complex nature of the distribution of $(\mathbf{I}-\mathbf{\Gamma}_{t+1})^{-1}$. However, the forecast distribution can be easily evaluated by simulation.

5 Real Data Example: Forecasting of S&P 500 Stocks

We run sequential learning and day ahead-forecasting on 400 members of the S&P 500. We investigate the effect of our Variational Bayes strategy on the forecasting performance of the SGDLM. The analysis is based on 2044 daily observations from January 2006 through October 2013.

We evaluate the forecast performance by the distribution quantiles of the realized returns of their respective forecast distributions (Figure 1). Ideally, the distribution of the quantiles should be uniform.



(a) With Variational Bayes

(b) No Variational Bayes

Figure 1: Distribution of the quantiles of the realizations in the $t+1$ forecast distributions across all series $j = 1 : 400$ and all 2044 observations.

It is evident from Figure 1 that the Variational Bayes step improves forecasting performance significantly. An analysis of the coverage properties of forecast intervals confirms this verdict on the level of individual stocks (Table 1).

Forecast interval	99.0%	95.0%	90.0%	80.0%	50.0%	20.0%	10.0%
	Across All Series						
Coverage w/ VB	98.6%	**95.9%**	**92.8%**	**85.8%**	**59.8%**	**27.2%**	**14.3%**
Coverage w/o VB	**99.2%**	97.9%	96.3%	92.9%	76.7%	41.6%	23.2%
	Apple Inc						
Coverage w/ VB	98.4%	**95.8%**	**92.3%**	**86.3%**	**60.1%**	**25.9%**	**13.4%**
Coverage w/o VB	**99.3%**	98.0%	96.2%	92.8%	74.8%	39.6%	19.7%
	Bank of America Corp						
Coverage w/ VB	**98.4%**	**95.8%**	**92.8%**	**86.0%**	**61.4%**	**27.3%**	**14.0%**
Coverage w/o VB	**98.4%**	96.7%	95.1%	92.1%	80.2%	50.8%	30.5%
	General Electric Co						
Coverage w/ VB	98.4%	**94.5%**	**91.3%**	**84.4%**	**58.7%**	**25.9%**	**12.9%**
Coverage w/o VB	**99.2%**	97.9%	95.8%	92.7%	77.9%	44.6%	25.3%
	McDonald's Corp						
Coverage w/ VB	98.6%	**96.1%**	**93.0%**	**86.7%**	**59.0%**	**26.1%**	**13.2%**
Coverage w/o VB	**99.3%**	98.6%	96.7%	92.9%	73.7%	38.5%	19.8%
	Pfizer Inc						
Coverage w/ VB	**98.9%**	**95.6%**	**92.5%**	**85.5%**	**59.5%**	**26.7%**	**14.3%**
Coverage w/o VB	99.6%	97.8%	96.6%	93.2%	76.8%	39.9%	20.8%
	Starbucks Corp						
Coverage w/ VB	98.2%	**95.6%**	**92.5%**	**86.5%**	**59.8%**	**27.0%**	**13.9%**
Coverage w/o VB	**98.8%**	97.2%	95.5%	92.5%	75.1%	39.3%	21.6%

Table 1: Coverage of centered $t+1$ forecast intervals of individual stock returns averaged over all 2044 observations.

References

[1] Omar Aguilar and Mike West. Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3):338–357, 2000.

[2] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory.* Wiley, 1994.

[3] C. M. Carvalho and Mike West. Dynamic matrix-variate graphical models. *Bayesian Analysis*, 2:69–98, 2007.

[4] John Geweke and Garland Durham. Massively parallel sequential monte carlo for bayesian inference. 2011.

[5] H. Kahn and A. W. Marshall. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, 1953.

[6] A. Lee, C. Yau, M. Giles, A. Doucet, and C. Holmes. On the utility of graphics cards to perform massively parallel simulation with advanced monte carlo methods. *Journal of Computational & Graphical Statistics*, 19(4), 2010.

[7] Raquel Prado and Mike West. *Time Series: Modelling, Computation & Inference.* Chapman & Hall/CRC Press, 2010.

[8] J. M. Quintana and M. West. An analysis of international exchange rates using multivariate dlms. *The Statistician*, 36:275–281, 1987.

[9] M. A. Suchard, Q. Wang, C. Chan, J. Frelinger, A. J. Cron, and M. West. Understanding gpu programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of Computational and Graphical Statistics*, 19:419–438, 2010. PMC2945379.

[10] Mike West and Jeff Harrison. *Bayesian Forecasting & Dynamic Models.* Springer Verlag, 2nd edition, 1997.

[11] Zoey Yi Zhao and Mike West. Dynamic compositional regression modelling: Application in financial time series forecasting and portfolio decisions. *Preprint*, 2013.