# Multi-Process Models –
# An Application for the Construction of Financial Factor Models

**Kevin R. Keane** and **Jason J. Corso**
Computer Science and Engineering
University at Buffalo, The State University of New York
Buffalo, NY 14260

## Abstract

We present an unsupervised, comprehensive methodology for the construction of financial risk models. We offer qualitative comments on incremental functionality and quantitative measures of superior performance of component and mixture dynamic linear models relative to alternative models. We apply our methodology to a high dimensional stream of daily closing prices for approximately 7,000 US traded stocks, ADRs, and ETFs for the most recent 10 years. Our methodology automatically extracts an evolving set of explanatory time series from the data stream; maintains and updates parameter distributions for component dynamic linear models as the explanatory time series evolve; and, ultimately specifies time-varying asset specific mixture models. Our methodology utilizes a hierarchical Bayesian approach for the specification of component model parameter distributions and for the specification of the mixing weights in the final model. Our approach is insensitive to the exact number of factors, and "effectively" sparse, as irrelevant factors (time series of pure noise) yield posterior parameter distributions with high density around zero. The statistical models obtained serve a variety of purposes, including: outlier detection; portfolio construction; and risk forecasting.

## 1 INTRODUCTION

We propose a time varying Bayesian statistical model for individual stock returns depicted in Figure 1. Our goal is to accurately model the high dimensional probability distribution underlying stock returns. We demonstrate success by constructing better performing investment portfolios, where the performance goal is to minimize the standard deviation of returns. Investment professionals seek to minimize the variation in investment outcomes because doing so results in higher economic utility for their clients. To illustrate this point, consider which of two pension scenarios with equal expected value is preferred: one that pays 80% salary with probability 1; or, the other that pays 120% salary with probability 1/2 (if things "go well") and pays 40% salary with probability 1/2 (if things "go poorly"). The economic concept of declining marginal utility, expressed mathematically with concave utility functions $\mathrm{U}(\mathrm{E}(x)) > \mathrm{E}(\mathrm{U}(x))$, implies that given investment scenarios with equal expected return, individuals prefer the scenario with lowest variation. Portfolio managers are concerned with generating acceptable returns with minimal risk; and, risk managers monitor the portfolio managers, verifying financial risks remain within authorized limits. Risk models, defining the $n \times n$ asset covariance matrix $\mathbf{\Sigma}$ and precision matrix $\mathbf{\Sigma}^{-1}$, are used by portfolio managers in conjunction with expected return vectors $\boldsymbol{\alpha}$ to *construct* optimal portfolios weights $\mathbf{\Sigma}^{-1}\alpha$; and, are used by risk managers given portfolio weights $\mathbf{w}$ to *forecast* portfolio variance $\mathbf{w}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{w}$.

We describe the construction of a set of Bayesian switching state-space models and qualitatively analyze the on-line behavior of the various component models and mixture models. We focus our discussion on the qualitative aspects then conclude by providing quantitative measures of our model's superior performance relative to competing models. We look at the tradeoff of adaption rate and stability of parameter estimates, evaluating model responsiveness with both synthetic and real data series. We show that dynamic linear models are robust [1] with respect to pure noise explanatory variables, appropriately generating param-

[1] We use the term *robust* to describe a desirable trait where an estimation method is stable in the presence of outlier observations and irrelevant explanatory variables (noise).

eter distributions very concentrated around zero. We comment on the behavior of component models relative to the mixture consensus. We illustrate component and mixture response to outlier observations. In periods with ambiguous posterior model probabilities, we describe the diffusive impact to the mixture distribution; and, we note surprisingly distinct behavior when an outlier component is selected with near certainty. We find unexpectedly similar updating behavior across a range of component models, bringing into question the necessity of more than two components. We inspect the impact of implementing intervention in the estimation process by greatly inflating the variance of a stock's posterior distribution subsequent to a merger event. The intervention is shown to result in extremely rapid convergence to new dynamics, while the same model without intervention is shown to maintain bias for an unacceptably long period. Lastly, we compare our two component mixture model against several alternatives. Analyzing results in Table 1, we show the positive impact of regularization provided by the Bayesian framework relative to PCA, and further improvement of the mixture model as compared to the single process model.

## 2 BACKGROUND

### 2.1 RISK MODELS

High dimensional statistical models, central to modern portfolio management, present significant technical challenge in their construction. There is extensive literature on the topic of constructing *factor models* or *risk models* as they are interchangeably known by practitioners. Consider a matrix of observations

$$\mathbf{X} = \begin{bmatrix} r_{1,1} & \cdots & r_{1,t} \\ \vdots & \ddots & \vdots \\ r_{n,1} & \cdots & r_{n,t} \end{bmatrix} \quad , \tag{1}$$

representing log price returns

$$r_{i,j} = \log\left(\frac{p_{i,j}}{p_{i,j-1}}\right) \quad , \tag{2}$$

for assets $i \in 1 \ldots n$, trading days $j \in 1 \ldots t$, and end of day prices $p_{i,j}$. The general approach to financial risk modeling specifies the covariance of asset returns as a structured matrix. A factor model with $p$ explanatory time series is specified for the $n \times t$ matrix $\mathbf{X}$ with an $n \times p$ matrix of common factor loadings $\mathbf{L}$, a $p \times t$ matrix of common factor returns time series $\mathbf{F}$, and an $n \times t$ matrix of residual error time series $\boldsymbol{\epsilon}$:

$$\mathbf{X} = \mathbf{LF} + \boldsymbol{\epsilon} \quad . \tag{3}$$

An orthogonal factor model (Johnson and Wichern, 1998, Ch. 9) implies diagonal covariance matrices

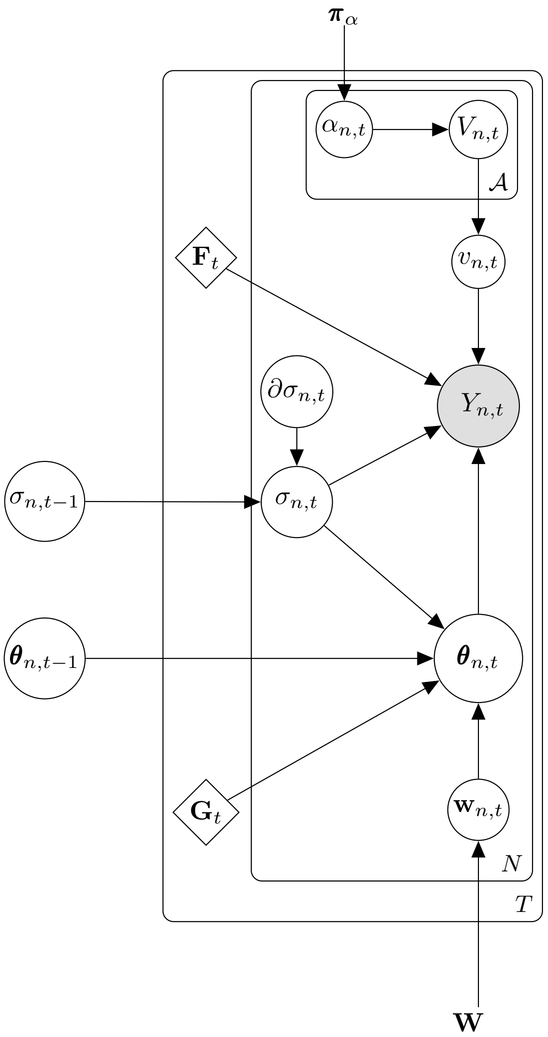


Figure 1: Our final two process Bayesian switching state-space model for log price returns $Y_{n,t} = \log(p_{n,t}/p_{n,t-1})$ for asset $n$ in time period $t$. We specify the prior probabilities $\pi_\alpha$ of the switch variable $\alpha_{n,t}$ that controls the observation variance: $V_{n,t} = 1$ if $\alpha_{n,t} = \{\textit{regular model}\}$ and $V_{n,t} = 100$ if $\alpha_{n,t} = \{\textit{outlier model}\}$. In our final model, we specify the evolution variance $\mathbf{W}$. Other variables are obtained or inferred from the data in an unsupervised manner. The return of an individual asset is modeled as a time varying regression, with regression coefficient vector $\boldsymbol{\theta}_{n,t}$, common factor explanatory vector $\mathbf{F}_t$, and noise: $Y_{n,t} = \boldsymbol{\theta}^{\mathsf{T}}_{n,t}\mathbf{F}_t + \sigma_{n,t}v_{n,t}$, $v_{n,t} \sim \mathrm{N}(0, V_{n,t})$. The regression coefficients are a hidden state vector, evolving as a Markov chain: $\boldsymbol{\theta}_{n,t} = \mathbf{G}_t\boldsymbol{\theta}_{n,t-1} + \sigma_{n,t}\mathbf{w}_{n,t}$, $\mathbf{w}_{n,t} \sim \mathrm{N}(\mathbf{0}, \mathbf{W})$. $\mathbf{G}_t$ captures rotation and scaling of the explanatory vector $\mathbf{F}_t$ over time, permitting computation of prior distributions for $\boldsymbol{\theta}_{n,t}$ from posterior distributions for $\boldsymbol{\theta}_{n,t-1}$. $\sigma_{n,t}$ is an asset and time specific scale factor applied to both the observation noise $v_{n,t}$ and the state evolution noise $\mathbf{w}_{n,t}$.

$\mathrm{Cov}(\mathbf{F}) = \mathbf{I}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}$. In an orthogonal factor model, the covariance of the observation matrix $\mathbf{X}$ is simply:

$$\mathrm{Cov}(\mathbf{X}) = \mathbf{L}\mathbf{L}^{\mathsf{T}} + \boldsymbol{\Psi} \quad . \tag{4}$$

By construction, risk models based on principal component and singular value decomposition (Wall et al., 2003), including ours, possess this simple structure of orthogonal common factors and residual errors.

## 2.2 COMMON FACTORS

In pursing an unsupervised methodology, we utilize an SVD based approach to identifying characteristic time series. SVD identifies common variation, both row-wise and column-wise, in a matrix of data (Wall et al., 2003). SVD decomposes a rectangular matrix, such as the returns $\mathbf{X}$, into two orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$; and, a diagonal matrix $\mathbf{D}$:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{T}} \quad . \tag{5}$$

Given the format of the $n \times t$ returns matrix $\mathbf{X}$ in Equation 3, the mutually orthogonal, unit-length right singular vectors comprising matrix $\mathbf{V}$ that represent common variation across the columns ($t$ trading days) are the *characteristic time series*; the mutually orthogonal, unit-length left singular vectors comprising matrix $\mathbf{U}$ that represent common variation across rows ($n$ assets) are the *characteristic portfolios*; and, the diagonal singular values matrix $\mathbf{D}$ captures scale. The entries of $\mathbf{D}$ are ordered by magnitude, therefore a $p$-factor model would use the first $p$-rows of $\mathbf{V}^{\mathsf{T}}$ for the factor return time series.

We exploit the fact that SVD methodically extracts common variation, ordered by magnitude. The intuition is that "pervasive" sources of return important to modeling portfolio level return characteristics will be reliably captured in the first several factors. Points of concern for some practitioners with regards to the use of SVD or PCA include:

1. the factors are not readily identifiable (to the human analyst);
2. the factors can be permuted in order and sign for data samples adjacent in time; and
3. it's not clear how many factors to "keep".

With respect to the first concern of (human) identifiability, we reiterate our goal is an unsupervised process yielding a high dimensional statistical model with adequate explanatory power as opposed to semantically meaningful groupings. Examination of assets with significant weight in the characteristic portfolios typically yields meaningful portfolio themes (Johnson and Wichern, 1998, Ex. 8.5). With respect to the second concern, the rotation and scaling of factors in different sample periods, our application incorporates the method of (Keane and Corso, 2012) to identify these rotations and maintain parameter distributions in the presence of rotation and scaling. Although much discussion surrounds the third concern, the identification of the "correct" number of factors (Roll and Ross, 1980; Trzcinka, 1986; Connor and Korajczyk, 1993; Onatski, 2010), we find the regularization provided by Bayesian dynamic linear models results in regression coefficients densely centered around zero for factors that are pure noise. The functioning of our process in this regard is analogous to regularized least squares (RLS) (Bishop, 2006, Ch. 3) and the ability of RLS to successfully estimate regression coefficients when confronted with a large number of candidate explanatory variables.

## 2.3 DYNAMIC LINEAR MODELS

The Bayesian dynamic linear model (DLM) framework elegantly addresses our need to process *streams* of data. DLMs are state space models very similar to Kalman filters (Kalman, 1960) and linear dynamical systems (Bishop, 2006, Ch. 13). We summarize the matrix variate notation and results from (West and Harrison, 1997, Ch. 16.4). Define $t$ the time index, $p$ the number of common factors, and $n$ the number of assets. The observations $\mathbf{Y}_t$ are generated by matrix variate dynamic linear models characterized by four time varying parameters, $\left\{\mathbf{F}_t, \mathbf{G}_t, V_t, \mathbf{W}_t\right\}$ that define the observation and state evolution equations. We now define and comment on the DLM parameters as they pertain to our application:

- $\mathbf{Y}_t = \left[Y_{t,1}, \ldots, Y_{t,n}\right]^{\mathsf{T}}$, log price returns at time $t$, common to all component DLMs;
- $\mathbf{F}_t$ a $p \times 1$ dynamic regression vector, factor returns at time $t$, common to all component DLMs;
- $\mathbf{G}_t$ a $p \times p$ state evolution matrix, accounts for rotation and scaling of factor return time series at time $t$, common to all component DLMs;
- $V_t$ an observational variance scalar, individually specified for each component DLM, greatly inflated in the DLMs generating "outliers" at time $t$;
- $\mathbf{W}_t$ an evolution variance matrix, individually specified for each component DLM, controls rate of change in factor loadings at time $t$;
- $\boldsymbol{\Sigma}_t = \begin{bmatrix} \sigma^2_{t,1} & & \\ & \ddots & \\ & & \sigma^2_{t,n} \end{bmatrix}$, unknown diagonal ma-

trix composed of $n$-asset specific variance scales at time $t$;

- $\boldsymbol{\nu}_t$, $n \times 1$ vector of unknown observation errors at time $t$;
- $\boldsymbol{\Theta}_t = \left[\theta_{t,1}, \ldots, \theta_{t,n}\right]$, a $p \times n$ unknown state matrix whose columns are factor loadings of individual assets on common factor returns at time $t$; and,
- $\boldsymbol{\Omega}_t = \left[\omega_{t,1}, \ldots, \omega_{t,n}\right]$, a $p \times n$ unknown evolution errors matrix applied to the state matrix at time $t$.

We specify our model variances in scale free form (West and Harrison, 1997, Ch. 4.5), implying multiplication by asset specific scales $\sigma^2_{t,i}$ in univariate DLMs: $\mathbf{C}_t = \sigma^2_{t,i}\mathbf{C}^*_t$, $V_t = \sigma^2_{t,i}V^*_t$, and $\mathbf{W}_t = \sigma^2_{t,i}\mathbf{W}^*_t$. When using matrix variate notation, $\boldsymbol{\Sigma}_t$ is a right variance parameter, discussed below, scaling the $n$ columns of the matrix on which it operates. When we specify models in § 3.5, we will specify scale free parameters $V^*_t \in \left\{1, 100\right\}$ and $\mathbf{W}^*_t \in \left\{.00001, .001, .1\right\}$. For simplicity, we shall omit scale free notation.

The observation equation is:

$$\mathbf{Y}_t = \boldsymbol{\Theta}_t^\top \mathbf{F}_t + \boldsymbol{\nu}_t, \quad \boldsymbol{\nu}_t \sim \mathrm{N}[\mathbf{0}, V_t\boldsymbol{\Sigma}_t] \quad . \qquad (6)$$

The distribution of the observation errors $\boldsymbol{\nu}_t$ is multivariate normal, with mean vector $\mathbf{0}$ and unknown variance $V_t\boldsymbol{\Sigma}_t$.

The state evolution is:

$$\boldsymbol{\Theta}_t = \mathbf{G}_t\boldsymbol{\Theta}_{t-1} + \boldsymbol{\Omega}_t, \quad \boldsymbol{\Omega}_t \sim \mathrm{N}[\mathbf{0}, \mathbf{W}_t, \boldsymbol{\Sigma}_t] \quad . \qquad (7)$$

The distribution of the state matrix evolution errors $\boldsymbol{\Omega}_t$ is *matrix normal* (Dawid, 1981), with mean matrix $\mathbf{0}$, left variance matrix $\mathbf{W}_t$ controlling variation across rows (factors) of $\boldsymbol{\Theta}_t$, and right variance matrix $\boldsymbol{\Sigma}_t$ controlling variation across columns (assets) of $\boldsymbol{\Theta}_t$. As our implementation uses diagonal matrices for both $\mathbf{W}_t$ and $\boldsymbol{\Sigma}_t$, we implicitly assume independence in the evolution of factor loadings across the factors $i \in 1 \ldots p$ and across the assets $j \in 1 \ldots n$.

Mapping factor model notation of (Johnson and Wichern, 1998) in Equation 3 to DLM notation of (West and Harrison, 1997) in Equation 6: $\mathbf{X} \rightarrow \left[\mathbf{Y}_1 \ldots \mathbf{Y}_t\right]$; $\mathbf{L} \rightarrow \boldsymbol{\Theta}_t^\top$; $\mathbf{F} \rightarrow \left[\mathbf{F}_1 \ldots \mathbf{F}_t\right]$; and, $\boldsymbol{\epsilon} \rightarrow \left[\boldsymbol{\nu}_1 \ldots \boldsymbol{\nu}_t\right]$. The crucial change in perspective involves the regression coefficients, $\mathbf{L} \rightarrow \boldsymbol{\Theta}_t^\top$. Where as the other matrices in Equation 3 are simply collections of columns present in Equation 6, the static regression coefficients $\mathbf{L}$ now evolve with time in Equation 7 as $\boldsymbol{\Theta}_t^\top$.

As typical with a Bayesian approach, our process begins with a prior distribution reflecting our belief about the unknown state parameter matrix $\boldsymbol{\Theta}_0$ and the unknown variance scale matrix $\boldsymbol{\Sigma}_0$ before data arrives. Our initial belief is expressed as a *matrix normal/inverse Wishart* distribution:

$$(\boldsymbol{\Theta}_0, \boldsymbol{\Sigma}_0) \sim \mathrm{NW}^{-1}_{\delta_0}[\mathbf{m}_0, \mathbf{C}_0, \mathbf{S}_0] \quad , \qquad (8)$$

with mean matrix $\mathbf{m}_0$, left variance matrix $\mathbf{C}_0$, right variance matrix $\mathbf{S}_0$, and degrees of freedom $\delta_0$. We allow our estimate of observational variance to vary over time by decaying our sample variance degrees of freedom parameter $\delta_{t-1}$ immediately before computation of the prior distribution Equation 10.

The marginal distribution for the state matrix $\boldsymbol{\Theta}_0$ is a *matrix T* distribution:

$$\boldsymbol{\Theta}_0 \sim \mathrm{T}_{\delta_0}[\mathbf{m}_0, \mathbf{C}_0, \mathbf{S}_0] \quad . \qquad (9)$$

Let $\mathbf{D}_t = \left[\mathbf{Y}_t \ldots \mathbf{Y}_0\right]$ refer to the information available subsequent to observing $\mathbf{Y}_t$. The conjugate parameter distributions are updated as follows.

Prior distribution at $t$:

$$(\boldsymbol{\Theta}_t, \boldsymbol{\Sigma}_t|\mathbf{D}_{t-1}) \sim \mathrm{NW}^{-1}_{\delta_{t-1}}[\mathbf{a}_t, \mathbf{R}_t, \mathbf{S}_{t-1}] \quad , \qquad (10)$$

where $\mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1}$ and $\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^\top + \mathbf{W}_t$.

Forecast distribution at $t$ given the dynamic regression vector $\mathbf{F}_t$:

$$(\mathbf{Y}_t|\mathbf{D}_{t-1}) \sim \mathrm{T}_{\delta_{t-1}}[\mathbf{f}_t, Q_t\mathbf{S}_{t-1}] \quad , \qquad (11)$$

where $\mathbf{f}_t = \mathbf{a}_t^\top\mathbf{F}_t$ and $Q_t = \left\{V_t + \mathbf{F}_t^\top\mathbf{R}_t\mathbf{F}_t\right\}$.

In our application, $\mathbf{F}_t$ is not available until $\mathbf{Y}_t$ is observed. Therefore, we accommodate random regression vectors $\mathbf{F}_t$ (Wang et al., 2011, § 7). Define $\mu_{\mathbf{F}_t} = \mathrm{E}(\mathbf{F}_t|\mathbf{D}_{t-1})$ and $\boldsymbol{\Sigma}_{\mathbf{F}_t} = \mathrm{Cov}(\mathbf{F}_t|\mathbf{D}_{t-1})$. The forecast distribution with $\mathbf{F}_t$ unknown is $(\mathbf{Y}_t|\mathbf{D}_{t-1}) \sim \mathrm{T}_{\delta_{t-1}}[\hat{\mathbf{f}}_t, \hat{Q}_t\mathbf{S}_{t-1}]$, where the moment parameters of the multivariate T forecast distribution are now $\hat{\mathbf{f}}_t = \mathbf{a}_t^\top\mu_{\mathbf{F}_t}$ and $\hat{Q}_t\mathbf{S}_{t-1} = \left\{V_t + \mu_{\mathbf{F}_t}^\top\mathbf{R}_t\mu_{\mathbf{F}_t} + \mathrm{tr}\left(\mathbf{R}_t\boldsymbol{\Sigma}_{\mathbf{F}_t}\right)\right\}\mathbf{S}_{t-1} + \mathbf{a}_t^\top\boldsymbol{\Sigma}_{\mathbf{F}_t}\mathbf{a}_t$.

Posterior distribution at $t$:

$$(\boldsymbol{\Theta}_t, \boldsymbol{\Sigma}_t|\mathbf{D}_t) \sim \mathrm{NW}^{-1}_{\delta_t}[\mathbf{m}_t, \mathbf{C}_t, \mathbf{S}_t] \quad , \qquad (12)$$

with $\mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t\mathbf{e}_t^\top$, $\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t\mathbf{A}_t^\top Q_t$, $\delta_t = \delta_{t-1} + 1$ and $\mathbf{S}_t = \delta_t^{-1}\left[\delta_{t-1}\mathbf{S}_{t-1} + \mathbf{e}_t\mathbf{e}_t^\top/Q_t\right]$ where $\mathbf{A}_t = \mathbf{R}_t\mathbf{F}_t/Q_t$ and $\mathbf{e}_t = \mathbf{Y}_t - \mathbf{f}_t$.

### 2.4 UNIVARIATE DLMS

In § 2.3, we summarized results for matrix variate DLMs. Setting the number of assets $n = 1$, results for univariate DLMs immediately follow.

## 2.5 MULTI-PROCESS MODELS

(West and Harrison, 1997, Ch. 12) define multi-process models composed of component DLMs. Consider a set of DLMs $\mathcal{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$. Let $\alpha_t$ reference the component DLM realized at time $t$, $\mathcal{A}_{\alpha_t} \in \mathcal{A}$. If the observations $Y_t$ are generated with one unknown DLM $\alpha_t = \alpha$ for all time, the observations are said to follow a *multi-process, class I model.* If at different times $s \neq t$, the observations are generated by distinct DLMs $\alpha_s \neq \alpha_t$, the observations are said to follow a *multi-process, class II model.* In an unsupervised modeling process, we need to accommodate the arrival of both typical and outlier observations. We accomplish this with a multi-process class II model, where various component DLMs are appropriate for various subsets of the observations. We assume fixed model selection probabilities, $\pi_t(\mathcal{A}_\alpha) = \pi(\mathcal{A}_\alpha)$.

With class II models, there are $|\mathcal{A}|^t$ potential model histories for *each* asset. We avoid this explosion in model sequences by considering only two periods, $t-1$ and $t$, thereby limiting distinct model sequences in our mixtures to $|\mathcal{A}|^2$. As each asset has its own history, no longer sharing common scale free posterior variance $\mathbf{C}_t$, our mixture models are asset specific, forcing the use of univariate component DLMs. Parameters $1 \times 1$ in univariate DLMs are now displayed with scalar notation. To avoid clutter, we omit implicit asset subscripts.

Inference with multi-process models is based upon manipulation of various model probabilities: the posterior model probabilities for the last model $p_{t-1}(\alpha_{t-1})$; the prior model probabilities for the current model $\pi(\alpha_t)$; and, the model sequence likelihoods $p(Y_t|\alpha_t, \alpha_{t-1}, D_{t-1})$. Posterior model sequence probabilities for current model $\alpha_t$ and last model $\alpha_{t-1}$ upon observing $Y_t$ are:

$$\begin{aligned} p_t(\alpha_t, \alpha_{t-1}) &= \Pr[\alpha_t, \alpha_{t-1}|D_t] \\ &\propto p_{t-1}(\alpha_{t-1})\pi(\alpha_t)p(Y_t|\alpha_t, \alpha_{t-1}, D_{t-1}) \quad . \end{aligned} \tag{13}$$

The unconditional posterior parameter distributions are computed as mixtures of the $|\mathcal{A}|^2$ component DLM sequences

$$p(\boldsymbol{\theta}_t|D_t) = \sum_{\alpha_t=1}^{k} \sum_{\alpha_{t-1}=1}^{k} p_t(\boldsymbol{\theta}_t|\alpha_t, \alpha_{t-1}, D_t)p_t(\alpha_t, \alpha_{t-1}) \; . \tag{14}$$

The posterior model probabilities are

$$p_t(\alpha_t) = \Pr[\alpha_t|D_t] = \sum_{\alpha_{t-1}=1}^{k} p_t(\alpha_t, \alpha_{t-1}) \quad . \tag{15}$$

The posterior probabilities for the last model $\alpha_{t-1}$ given the current model $\alpha_t$ and information $D_t$ are

$$\Pr[\alpha_{t-1}|\alpha_t, D_t] = \frac{p_t(\alpha_t, \alpha_{t-1})}{p_t(\alpha_t)} \quad . \tag{16}$$

After each time step, the posterior mixture distribution for each component DLM is approximated with an analytic distribution using the methodology described in (West and Harrison, 1997, Ch. 12.3.4). The Kullback-Leibler directed divergence between the approximation and the mixture is minimized in the parameters: $\mathbf{m}_t(\alpha_t)$, $\mathbf{C}_t(\alpha_t)$, $S_t(\alpha_t)$, and $\delta_t(\alpha_t)$. Let $S_t(\alpha_t, \alpha_{t-1})$ refer to the variance scale estimate obtained with the DLM sequence $\alpha_{t-1}$, $\alpha_t$. The parameters of the approximating distributions are as follows. The variance scale estimates $S_t(\alpha_t)$ are:

$$S_t(\alpha_t)^{-1} = \frac{1}{p_t(\alpha_t)} \sum_{\alpha_{t-1}=1}^{k} \frac{p_t(\alpha_t, \alpha_{t-1})}{S_t(\alpha_t, \alpha_{t-1})} \quad . \tag{17}$$

The weights for computing the moments of the KL divergence minimizing approximation to the posterior distribution are:

$$p_t^*(\alpha_{t-1}) = \frac{S_t(\alpha_t)}{p_t(\alpha_t)} \frac{p_t(\alpha_t, \alpha_{t-1})}{S_t(\alpha_t, \alpha_{t-1})} \quad . \tag{18}$$

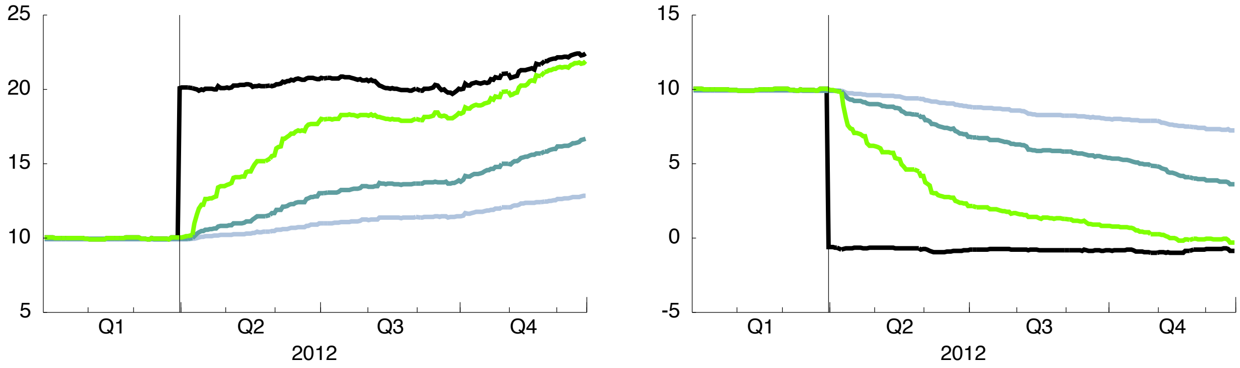
The mean vector $\mathbf{m}_t(\alpha_t)$ for DLM $\alpha_t$ is:

$$\mathbf{m}_t(\alpha_t) = \sum_{\alpha_{t-1}=1}^{k} p_t^*(\alpha_{t-1})\mathbf{m}_t(\alpha_t, \alpha_{t-1}) \quad . \tag{19}$$

The variance matrix $\mathbf{C}_t(\alpha_t)$ for DLM $\alpha_t$ is:

$$\begin{aligned} \mathbf{C}_t(\alpha_t) = \sum_{\alpha_{t-1}=1}^{k} p_t^*(\alpha_{t-1}) \; \Big\{ \; & \mathbf{C}_t(\alpha_t, \alpha_{t-1}) + \\ & [\mathbf{m}_t(\alpha_t) - \mathbf{m}_t(\alpha_t, \alpha_{t-1})] \quad \times \\ & [\mathbf{m}_t(\alpha_t) - \mathbf{m}_t(\alpha_t, \alpha_{t-1})]^{\mathsf{T}} \; \Big\} \quad . \end{aligned} \tag{20}$$

The degrees of freedom parameter, $\delta_t(\alpha_t)$ is important to the KL minimization. Intuitively, if the component DLMs are in discord, the resulting mixture may be described as “fat-tailed”, and the precision of the unknown variance scale parameter reduced. We compute $\delta_t(\alpha_t)$ using the procedure described by (West and Harrison, 1997, Ex. 12.7), with a further correction term. The approximation West and Harrison utilize, based upon an algorithm for computing the digamma function $\gamma(x)$ discussed in (Bernardo, 1976), is appropriate when $x \to \infty$. However, when estimating the *reciprocal* of the KL minimizing $\delta_t(\alpha_t)$, we find the error in the approximation remains rather constant, and we apply a correction to eliminate this constant.



(a) Synthetic scenario 1: abrupt increase in factor loading $\theta_t$, units are percent per annum. Observations $Y_t$ are synthesized returns using SPY (S&P 500 ETF) until March 30, 2012; and, 2×SPY thereafter. Note "true" factor loading doubles from approximately 10 to 20.

(b) Synthetic scenario 2: abrupt decrease in factor loading $\theta_t$, units are percent per annum. Observations $Y_t$ are synthesized returns using SPY until March 30, 2012; and, AGG (Barclays Aggregate Bond ETF) thereafter. Note "true" factor loading drops from approximately 10 to 0.

Figure 2: Unmonitored mixture models responding to abrupt change. Black line is true value of latent state variable $\theta_{1,t}$. Other lines are the posterior mean $\mathbf{m}_{1,t}$ (first common factor loading) obtained with three different mixture models discussed in § 3.5. None of the three above models responded quickly enough to the dramatic change in dynamics. A method of intervention to improve responsiveness is discussed in § 4.5.

# 3 APPLICATION DESIGN

## 3.1 END USERS

Our application enables a proprietary trading group at a financial services firm to better understand the aggregate behavior of stock portfolios. The models provide a statistical framework required for constructing portfolios, assessing portfolio risk, and clustering assets. The assets of primary interest are the common shares of the largest 1000 - 2000 companies in the US stock market. The group focuses on larger companies because the liquidity of larger stocks generally makes them cheaper and easier to transact. The group's strategies include short selling: borrowing stock, selling borrowed shares; and, attempting to profit by repurchasing the shares at a lower price. Larger stocks are generally easier to borrow.

## 3.2 BIAS-VARIANCE TRADEOFF

In implementing our application, one of the first issues encountered is a machine learning classic, the bias-variance tradeoff [Ch. 2.9](Hastie et al., 2009). With respect to DLMs, a trade off is incurred in the effective number of observations as the evolution variance is varied. A model with greater evolution variance will generate parameter distributions with greater variance but lower bias. A model with lower evolution variance will generate parameter distributions with lower variance but greater bias. A relatively smooth, lethargic, slowly adapting model does not track evolving dynamics as quickly as a rapidly adapting model; on the other hand, the quickly adapting model delivers a noisier sequence of parameter distributions. Outside our applied context, the loss function might be specified as squared error or absolute error. In the context of a risk model, the loss function should consider a portfolio manager's cost of over-trading due to a model adapting excessively (variance); as well a risk manager's problems arising in a system adapting too slowly (bias). The appropriate loss function depends critically on the intended end use. A quantitative trader constructing portfolios with quadratic optimization tends to magnify errors in a model, as the optimization process responds dynamically to parameter estimates (Muller, 1993). In contrast, a firm-wide risk manager, who typically evaluates sums of individual asset exposures, but does not dynamically respond to individual asset risk attributes, may prefer less bias and more variance, as error in the factor loadings of one asset may be offset by error in another asset in the summation process. We construct a variety of mixtures along the bias-variance continuum as discussed in § 3.5 and as illustrated in Figure 2.

## 3.3 UNIVERSE OF ASSETS

We identify two universes of assets: a relatively narrow set that will be used to construct explanatory time series; and, an all inclusive set for which we will generate factor loading and residual volatility estimates. It is desirable that the assets used to construct the common factor returns trade frequently and with adequate liquidity to minimize pricing errors. We also avoid *survivor bias*, the methodological error of omitting companies no longer in existence at the time a historical analysis is performed. We eliminate this hazard by defining our common factor estimation universe using daily exchange traded fund (ETF) create / redeem portfolios for the last 10 years. Brokers *create* ETF shares (in units of 50,000 ETF shares) by delivering a defined portfolio of stock and cash in exchange for ETF shares; or, they *redeem* ETF shares and receive the defined portfolio of stock and cash. Given the create / redeem definitions that were used as the basis for large transactions during the historical period, the assets in an ETF portfolio represent an institutionally held, survivor bias free, tradeable universe. Its likely the component shares were readily available to borrow, as the component shares were held by custodian banks for the ETF shareholders. Our data base permits us to obtain data for surviving and extinct stocks. The ETF we select for our factor estimation universe is the Vanguard Total Stock Market ETF (VTI) (Vanguard Group, Inc., 2014). As of April 2014, including both mutual fund and ETF share classes, the Vanguard Total Stock Market fund size was approximately USD 330 billion. The VTI constituents closely approximates our desired universe, with the number of component stocks typically ranging from 1300 - 1800.

### 3.4 DATA PREPARATION

Data preparation involves constructing the artifacts demanded by § 2.3: $\mathbf{Y}_t$, $\mathbf{F}_t$, and $\mathbf{G}_t$. Each trading day, using price, dividend, and corporate action data for all 7000 - 8000 stocks, ADRs, and ETFs in our US pricing data base (MarketMap Analytic Platform, 2014), we construct dividend and split adjusted log price return observation vectors $\mathbf{Y}_t$. For stocks in the VTI ETF on that day, we construct a variance equalized historical returns matrix $\mathbf{r}_t = \left[\mathbf{Y}_{t-T+1} \ldots \mathbf{Y}_t\right] \hat{\mathbf{\Sigma}}_t^{-\frac{1}{2}}$ where $\hat{\mathbf{\Sigma}}_t$ is the diagonal matrix of sample variance for the period $t-T+1$ to $T$. Using (Keane and Corso, 2012, §3.c), we compute $\mathbf{F}_t$, from the first $p$ right singular vectors from a singular value decomposition of $\mathbf{r}_t$. As the vectors are unit length, and we desire unit variance per day, $\mathrm{Cov}(\mathbf{F}_t) = \mathbf{I}$, we scale the right singular vectors by $\sqrt{T}$. The scaled characteristic time series from adjacent data windows, $\mathbf{r}_{t-1} = \left[\mathbf{Y}_{t-T} \ldots \mathbf{Y}_{t-1}\right] \hat{\mathbf{\Sigma}}_{t-1}^{-\frac{1}{2}}$ and $\mathbf{r}_t = \left[\mathbf{Y}_{t-T+1} \ldots \mathbf{Y}_t\right] \hat{\mathbf{\Sigma}}_t^{-\frac{1}{2}}$ are then used to compute $\mathbf{G}_t$ as described in (Keane and Corso, 2012, §3.e):

$$\mathbf{G}_t = \left(\mathbf{F}_t \mathbf{F}_t^{\mathsf{T}}\right)^{-1} \mathbf{F}_t \mathbf{F}_{t-1}^{\mathsf{T}} \quad . \qquad (21)$$

We are a little flexible with the notation in Equation 21, where $\mathbf{F}_t$ and $\mathbf{F}_{t-1}$ are $p \times (T-1)$ sub-matrices representing time aligned subsets of two factor return matrices, the scaled right singular vectors obtained from the decomposition of $\mathbf{r}_{t-1}$ and $\mathbf{r}_t$. Elsewhere, viz. Equation 11, $\mathbf{F}_t$ refers to a $p \times 1$ dynamic regression vector, the right most column of the transposed and scaled right singular vectors, corresponding to the desired vector of common factor returns for day $t$.

### 3.5 MODEL COMPONENTS

The component DLMs in our mixture model share observations $\mathbf{Y}_t$, common factor scores $\mathbf{F}_t$, and state evolution matrices $\mathbf{G}_t$. The component DLMs are differentiated by the variance parameters: the observational variance scale $V_t$, and the evolution variance matrix $\mathbf{W}_t$. We construct a set of component DLMs following the approach of (West and Harrison, 1997, Ch. 12.4). For component DLMs that will accommodate "typical" observation variance, we set $V_t = 1$; for component DLMs that will accommodate outlier observations, we set $V_t = 100$. For the evolution variance, we similarly select a base rate of evolution $\mathbf{W}_t = .00001$; and, inflate $\mathbf{W}_t$ by a factor of 100 and $100^2$ to permit increasingly rapid changes in the factor loadings.

### 3.6 OUTPUT

The format of the output risk model will be a $n \times p$ factor loading matrix and a $n \times 1$ residual volatility vector. These $p+1$ numeric attributes for $n$ stocks are stored in various formats for subsequent use throughout the organization.

## 4 EVALUATION

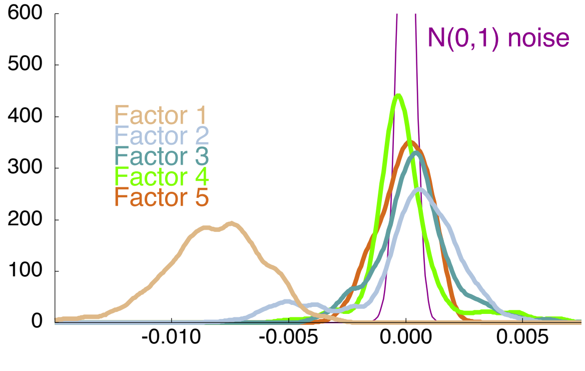
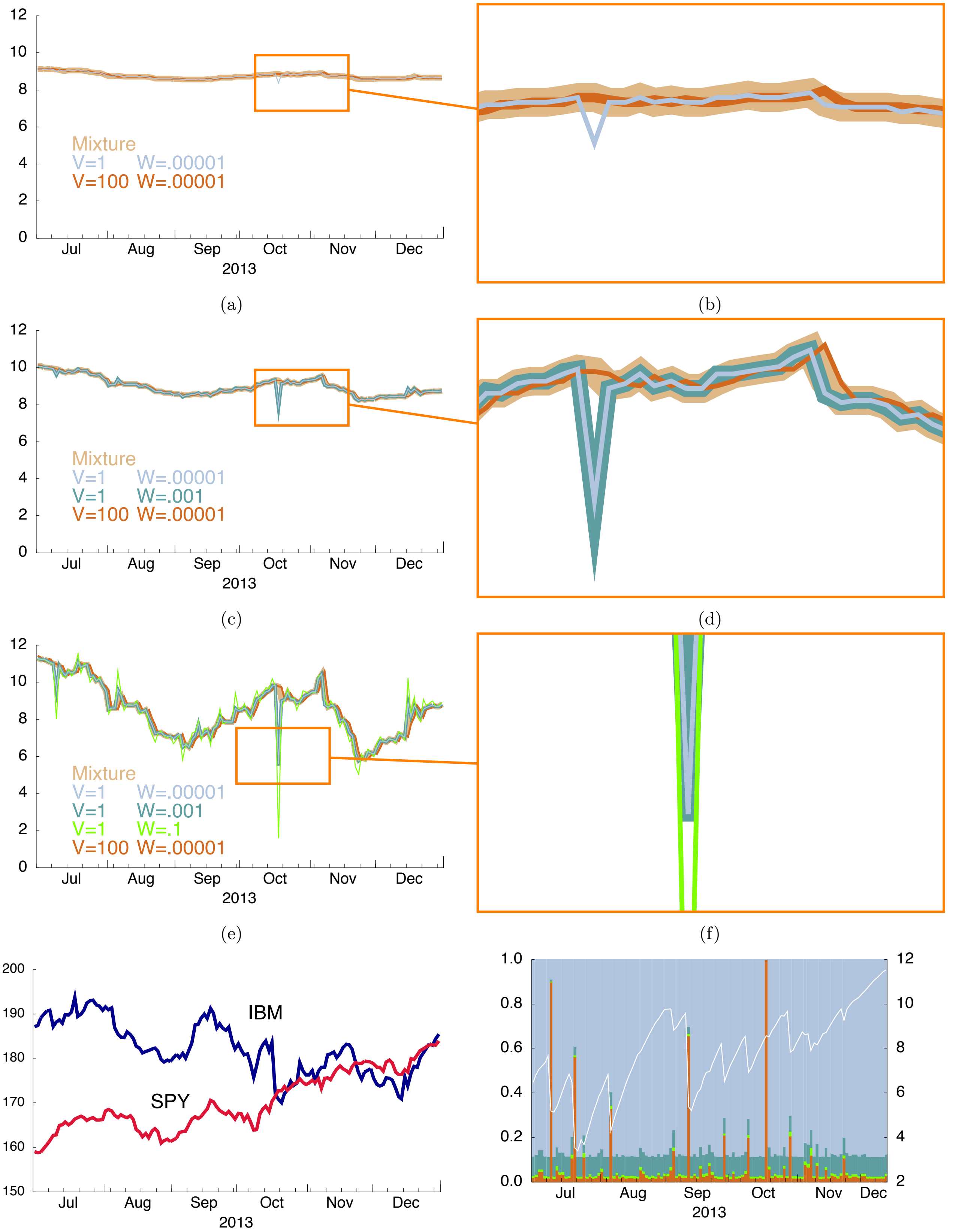


Figure 3: Distribution of factor loadings for the Vanguard Total Market Index portfolio on April 30, 2014. Left axis is density, $\int dx = 1$.

### 4.1 NUMBER OF FACTORS

A Bayesian DLM updated with an "explanatory" series of pure noise is expected to generate posterior



(a)

(b)

(c)

(d)

(e)

(f)

(g) Price histories for IBM and SPY, units are USD.

(h) Posterior component probabilities *(left)*, color key in (c); mixture model degrees of freedom $\delta_t$ *(right)*, white line.

Figure 4: Multi-Process Models. See § 4.3 for discussion. Figure 4(a),(c),(e) units are percent per annum.

regression parameter distributions densely surrounding zero given sufficient observations. Further, a linear combination of independent DLMs is a DLM (West and Harrison, 1997, *Principle of Superposition*, p. 188). We use these two points to justify the inclusion of a relatively large number of independent explanatory series. Given the regularization inherent in Bayesian DLMs, we believe the risk of omitting a common factor far exceeds the risk of including noise. In our application, we focus on aggregate (portfolio level) forecasts, therefore it is extremely important to identify common sources of variation that may appear insignificant at the individual asset level. (West and Harrison, 1997, Ch. 16.3) discuss in detail the hazard of omitted common factors in aggregate forecasts. In Figure 3, we show the distribution of factor loadings for the VTI constituents on April 30, 2014. The distribution of factor loadings for the first five common factors obtained from our two component mixture model are shown in comparison to the distribution of loadings on Gaussian noise, $F_t \sim \mathrm{N}[0, 1]$. Figure 3 is consistent with our viewpoint, note the high density of zero loadings for the noise series, and the relatively diffuse factor loadings for the first five common factor series extracted with SVD. The factor loadings on the noise series have a mean loading $\mu_m = 0$ bp[2], and a standard deviation of $\sigma = 2$ bp. In contrast, the loadings on the first factor have a mean loading of $\mu_m = -82$bp and a standard deviation of $\sigma_m = 21$bp.

### 4.2 MIXTURE DYNAMICS

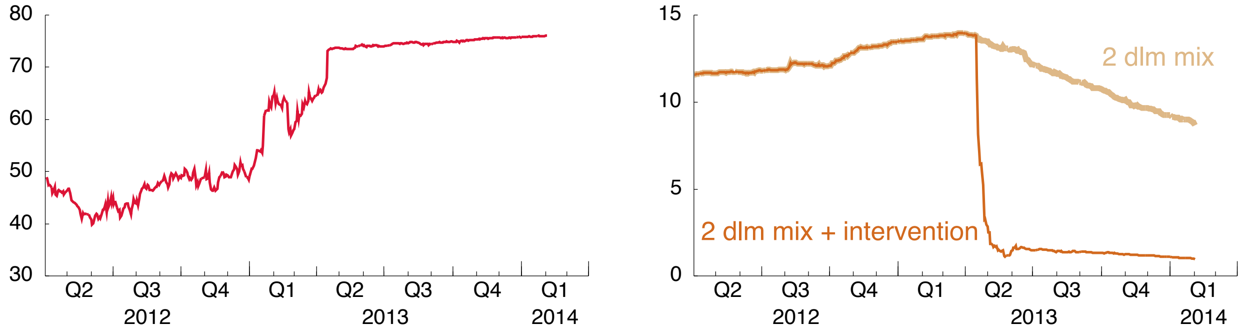
Figure 4 shows the price movement and model response for IBM during the second half of 2013. In Figure 4(g), the price histories for IBM and the S&P 500 ETF SPY are displayed. Note the sharp drop in IBM's price on October 17, 2013 corresponding to an earnings announcement. This event is helpful in understanding the interaction of components in our multi-process models. We construct three multi-process models, the simplest of which is a two component mixture (the "base model"), comprised of a standard component DLM to handle the majority of the observations $Y_t$, and an outlier component DLM. We specify common evolution variance $\mathbf{W}_t = .00001\ \mathbf{I}$; observation variance $\mathbf{V}_t = 1$ for the standard component; and observation variance $\mathbf{V}_t = 100$ for the outlier component. The base model and component estimates for the first factor loading $\mathbf{m}_{1,t}$ appear in Figure 4(a) and magnified in Figure 4(b). We specify a three component mixture (the "adaptive model") by adding a component DLM with inflated evolution variance $\mathbf{W}_t = .001\ \mathbf{I}$. The adaptive model and component estimates for the first factor loading $\mathbf{m}_{1,t}$ appear in Figure 4(c) and magnified in Figure 4(d). Finally, we specify a four component mixture (the "very adaptive model") by adding a component with further inflated evolution variance $\mathbf{W}_t = .1\ \mathbf{I}$. The very adaptive model and component estimates for the first factor loading $\mathbf{m}_{1,t}$ appear in Figure 4(e) and magnified in Figure 4(f). The posterior component model probabilities for the very adaptive model appear as a bar chart in Figure 4(h), where the bottom bar corresponds to the probability of the outlier component, the second bar corresponds to the very adaptive component DLM, the third bar corresponds to the adaptive component DLM, and the top bar corresponds to the base component DLM. We specified the fixed DLM selection (prior) probabilities as $\{.01543, .00887, .0887, .887\}$ for the components $\{$*outlier, very adaptive, adaptive, base*$\}$ respectively. Note several occurrences where the posterior probability of an outlier observation significantly exceeds the DLM selection probability. The white line in Figure 4(h) corresponds to the degrees of freedom parameter $\delta_t$ for the T-distribution that approximates the mixture model's posterior parameter distribution.

### 4.3 QUALITATIVE COMMENTS

To supplement the more analytically precise discussion in § 2.5, we make the following qualitative comments as to the interaction of the mixture components:

- the mixture in Figure 4(e) with larger evolution variance adapts faster; the mixture in Figure 4(a) with smaller evolution variance appears smoother;
- time $t$ component posteriors are 1-period departures from the $t-1$ consensus, see Figure 4(a), (b), (c), (d), and (e);
- an outlier component "ignores" current observations $Y_t$ and forecast error $|\mathrm{e}_t|$, responding to the $t-1$ posterior consensus $\mathbf{m}_{t-1}$, see Figure 4(b) and (d);
- in periods of noise, the other components return to the outlier component's estimate with 1-period lag, see left-hand side of Figure 4(b) and (d);
- in periods of level change, the outlier follows the other components' estimate with 1-period lag, see right-hand side of Figure 4(b) and (d);
- when the posterior probability of the outlier component spikes up, the degrees of freedom parameter $\delta_t$ usually drops, reducing the precision of the variance scale estimate $S_t$, see Figure 4(h);
- *however*, when the outlier component is selected with very high probability, there is no impact to $\delta_t$ as the observation is ignored, see October 17,

[2] A basis point (bp) is $10,000^{-1}$.



(a) Price history, units are USD.

(b) Posterior mean $\mathbf{m}_{1,t}$, units are percent per annum.

Figure 5: In a deal announced April 15, 2013, Life Technologies Corporation was acquired by Thermo Fisher Scientific Inc. for USD 14 billion. The change in price behavior is noticable in Figure 5a. The factor loading estimates for two mixture models are compared in Figure 5b. One mixture model is unmonitored, the other benefits from intervention subsequent to the news event. See discussion in § 4.5

2013 in Figure 4(h), noting that the white line does not drop when Pr{ *outlier* } ≈ 1;

- except for the very adaptive component, the response does not vary with $\mathbf{W}_t$, see Figure 4(d) and (f), where W=.001 and W=.00001 responses are nearly identical.

### 4.4 UNRESPONSIVENESS TO $\mathbf{W}_t$

The phenomenon we find most surprising is the insensitivity of the components to $\mathbf{W}_t$ below a certain threshold. Digging further into this phenomenon, in a mixture, the various component models view of the $t-1$ posterior parameter distribution are very similar. Thinking about univariate DLMs, and assuming for discussion $F = 1$ and $G = 1$, the magnitude of the adaptive scalar $A_t = R_t/Q_t = (C_{t-1} + W_t) \,/\, (C_{t-1} + W_t + V_t)$. When $W_t \ll C_{t-1}$, as describes our situation, $A_t \approx C_{t-1}/\,(C_{t-1} + V_t)$ as seen in Figure 4(d) and (f). Only when $W_t$ is significant relative to $C_{t-1}$ does response vary noticeably.

### 4.5 INTERVENTION

Our discussion of IBM focused on the mixture models' processing of unusual observations. We now explore an example where the data generating process changes abruptly, similar to our synthetic illustrations in Figure 2. In April 2013, the acquisition of Life Technologies Corporation by Thermo Fisher Scientific Inc. was announced. The stock's sensitivity to the market, as expressed in its first common factor loading, dropped abruptly, as shown in Figure 5. While our goal is an unsupervised estimation processes, the Bayesian DLM framework facilitates structured intervention when necessary. For one of the mixture models in Figure 5b, we intervene and inflate the prior parameter variance following the April 15th announcement, $\mathbf{R}_t^{++} = \mathbf{G}_t\,(\mathbf{C}_{t-1} + \mathbf{I})\,\mathbf{G}^{\top} + \mathbf{W}_t$, where the identity matrix reflects the increased uncertainty in the parameter distribution relative to the usual prior variance in Equation 10. When subsequent updates occur, the DLM with the inflated prior variance adapts to the new dynamics rapidly and satisfactorily.

Table 1: Risk Model Performance

| | Volatility | s.e. | *t*-stat |
|---|---|---|---|
| | | | |
| GMV portfolio | | | |
| PCA (1) | 4.62 | 0.07 | 40.48 |
| PCA (10) | 3.41 | 0.05 | 29.87 |
| DLM (10) | 2.17 | 0.03 | 6.82 |
| **Mixture (10)** | **1.96** | 0.03 | |
| | | | |
| MSR portfolio | | | |
| PCA (1) | 4.07 | 0.06 | 49.38 |
| PCA (10) | 2.06 | 0.03 | 28.84 |
| DLM (10) | 1.30 | 0.02 | 4.39 |
| **Mixture (10)** | **1.22** | 0.02 | |
| | | | |
| Cap Weight | 20.31 | 0.29 | |
| Equal Weight | 24.62 | 0.35 | |

### 4.6 RISK MODEL PERFORMANCE

To access the performance of our two component mixture model, we construct daily portfolios from the VTI universe for the most recent ten years, May 2004 to April 2014. The number of trading days during this period was 2,516. In Table 1, we report realized outof-sample volatility and the standard error (se) of the volatility measure for two strategies: global minimum variance (GMV); and, maximum Sharpe ratio (MSR) (Demey et al., 2010). We implement four risk models: 1-factor PCA; 10-factor PCA; 10-factor DLM; and, 10-factor 2-component mixture model. We permit short positions, and do not constrain position weights. The PCA models are constructed using (Connor and Korajczyk, 1988). The mixture model is the same model we presented earlier, with noise "factor ten" present. For each strategy, we report the *t*-statistics for the various models' realized volatility compared to the mixture model's realized volatility. For context on the ambient volatility of the ten year period, we provide realized volatility for two portfolios that are long only and do not use a risk model: the capitalization weighted portfolio; and the equal weighted portfolio.

## 5 CONCLUSION

The ability to integrate SVD with Bayesian methods allows our application to process large data streams in an unsupervised fashion. We demonstrate that a two component multi-process model achieved better reduction in volatility than alternative models. The two component model out-performed alternative models including a single process model. We find the robustness of Bayesian DLMs with respect to noise inputs of great practical value, allowing us to favor inclusion of factors, potentially capturing pervasive sources of common movement important to aggregate forecasting. The inclusion of an outlier model adds great functionality, delivering robustness to the estimation process. The insensitivity of the mixture models to multiple evolution variance values leads us to favor mixtures of just two components, a typical evolution variance value for both components, and an inflated observation variance in the outlier component. We would recommend generating several models of varying adaptiveness, evaluating the variance-bias tradeoff in light of a user's specific situation. We favor the judicious use of intervention for events such as mergers. We would like to explore using news feeds to systematically intervene for events known to impact an assets dynamics.

### References

J.M. Bernardo. Psi (digamma) function. *Applied Statistics*, 25(3):315–317, 1976.

C.M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.

G. Connor and R.A. Korajczyk. Risk and return in an equilibrium APT. *Journal of Financial Economics*, 21(2):255–289, 1988.

G. Connor and R.A. Korajczyk. A test for the number of factors in an approximate factor model. *The Journal of Finance*, 48(4):1263–1291, 1993.

A.P. Dawid. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265, 1981.

P. Demey, S. Maillard, and T. Roncalli. Risk based indexation. Technical report, Lyxor Asset Management, Paris, 2010.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*, volume 2. Springer, 2009.

R.A. Johnson and D.W. Wichern. *Applied multivariate statistical analysis*, volume 4. Prentice Hall, 1998.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

K.R. Keane and J.J. Corso. Maintaining prior distributions across evolving eigenspaces. In *International Conference on Machine Learning and Applications.* IEEE, 2012.

MarketMap Analytic Platform. *North American Pricing.* SunGard Data Systems, Inc., 2014.

P. Muller. Empirical tests of biases in equity portfolio optimization. In S.A. Zenios, editor, *Financial optimization.* Cambridge University Press, 1993.

A. Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, 2010.

R. Roll and S.A. Ross. An empirical investigation of the arbitrage pricing theory. *The Journal of Finance*, 35(5):1073–1103, 1980.

C. Trzcinka. On the number of factors in the arbitrage pricing model. *the Journal of Finance*, 41(2):347–368, 1986.

The Vanguard Group, Inc. Prospectus. Vanguard Total Stock Market ETF. https://www.vanguard.com, 2014.

M.E. Wall, A. Rechtsteiner, and L.M. Rocha. Singular value decomposition and principal component analysis. In D.P. Berrar, W. Dubitzky, and M. Granzow, editors, *A practical approach to microarray data analysis.* Springer, 2003.

H. Wang, C. Reeson, and C. Carvalho. Dynamic financial index models: Modeling conditional dependencies via graphs. *Bayesian Analysis*, 6(4):639–664, 2011.

M. West and J. Harrison. *Bayesian forecasting and dynamic models.* Springer Verlag, 1997.