DYNAMICALLY MIXING DYNAMIC LINEAR MODELS WITH APPLICATIONS IN FINANCE

Kevin R. Keane and Jason J. Corso
Department of Computer Science and Engineering
University at Buffalo, The State University of New York, Buffalo, NY, U.S.A.

Keywords: Bayesian inference, Dynamic linear models, Multi-process models, Statistical arbitrage.

Abstract: Time varying model parameters offer tremendous flexibility while requiring more sophisticated learning methods. We discuss on-line estimation of time varying DLM parameters by means of a dynamic mixture model composed of constant parameter DLMs. For time series with low signal-to-noise ratios, we propose a novel method of constructing model priors. We calculate model likelihoods by comparing forecast distributions with observed values. We utilize computationally efficient moment matching Gaussians to approximate exact mixtures of path dependent posterior densities. The effectiveness of our approach is illustrated by extracting insightful time varying parameters for an ETF returns model in a period spanning the 2008 financial crisis. We conclude by demonstrating the superior performance of time varying mixture models against constant parameter DLMs in a statistical arbitrage application.

1 BACKGROUND

1.1 Linear Models

Linear models are utilitarian work horses in many domains of application. A model's linear relationship between a *regression vector* F_t and an *observed response* Y_t is expressed through coefficients of a *regression parameter vector* θ. Allowing an *error of fit* term ε_t, a linear regression model takes the form:

$$Y = F^\mathsf{T}\theta + \varepsilon \quad , \tag{1}$$

where Y is a column vector of individual observations Y_t, F is a matrix with column vectors F_t corresponding to individual regression vectors, and ε a column vector of individual errors ε_t.

The vector Y and the matrix F are observed. The *ordinary least squares* ("OLS") estimate $\hat{\theta}$ of the regression parameter vector θ is (Johnson and Wichern, 2002):

$$\hat{\theta} = \left(FF^\mathsf{T}\right)^{-1} FY \quad . \tag{2}$$

1.2 Stock Returns Example

In modeling the returns of an individual stock, we might believe that a stock's return is roughly a linear function of market return, industry return, and stock specific return. This could be expressed as a linear model in the form of (1) as follows:

$$r = F^\mathsf{T}\theta + \varepsilon, \quad F = \begin{bmatrix} 1 \\ r_M \\ r_I \end{bmatrix}, \quad \theta = \begin{bmatrix} \alpha \\ \beta_M \\ \beta_I \end{bmatrix}, \tag{3}$$

where r represents the stock's return, r_M is the market return, r_I is the industry return, α is a stock specific return component, β_M is the sensitivity of the stock to market return, and β_I is the sensitivity of the stock to it's industry return.

1.3 Dynamic Linear Models

Ordinary least squares, as defined in (2), yields a single estimate $\hat{\theta}$ of the regression parameter vector θ for the entire data set. Problems arise with this framework if we don't have a *finite* data set, but rather an *infinite* data stream. We might expect θ, the coefficients of a linear relationship, to vary slightly over time $\theta_t \approx \theta_{t+1}$. This motivates the introduction of *dynamic linear models* (West and Harrison, 1997). DLMs are a generalized form, subsuming Kalman filters (Kalman et al., 1960), flexible least squares (Kalaba and Tesfatsion, 1996), linear dynamical systems (Minka, 1999; Bishop, 2006), and several time series methods — Holt's point predictor, exponentially weighted moving averages, Brown's exponentially weighted regres-

R. Keane K. and J. Corso J. (2012).
DYNAMICALLY MIXING DYNAMIC LINEAR MODELS WITH APPLICATIONS IN FINANCE.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 295-302
DOI: 10.5220/0003712602950302

sion, and Box-Jenkins autoregressive integrated moving average models (West and Harrison, 1997). The regime switching model in (Hamilton, 1994) may be expressed as a DLM, specifying an autoregressive model where evolution variance is zero except at times of regime change.

1.4 Contributions and Paper Structure

The remainder of the paper is organized as follows. In section §2, we introduce DLMs in further detail; discuss updating estimated model parameter distributions upon arrival of incremental data; show how forecast distributions and forecast errors may be used to evaluate candidate models; the generation of data given a DLM specification; inference as to which model was the likely generator of the observed data; and, a simple example of model inference using synthetic data with known parameters. Building upon this base, in section §3 multi-process mixture models are introduced. We report design challenges we tackled in implementing a mixture model for financial time series. In section §4, we introduce an alternative set of widely available financial time series permitting easier replication of the work in (Montana et al., 2009); and we provide an example of applying a mixture model to real world financial data, extracting insightful time varying estimates of variance in an ETF returns model during the recent financial crisis. In section §5, we augment the statistical arbitrage strategy proposed in (Montana et al., 2009) by incorporating a hedge that significantly improves strategy performance. We demonstrate that an on-line dynamic mixture model outperforms all statically parameterized DLMs. Further, we draw attention to the fact that the period of unusually large mispricing identified by our mixture model coincides with unusually high profitability for the statistical arbitrage strategy. In §6, we conclude.

2 DYNAMIC LINEAR MODELS

2.1 Specifying a DLM

In the framework of (West and Harrison, 1997), a dynamic linear model is specified by its parameter quadruple $\{F_t, G, V, W\}$. DLMs are controlled by two key equations. One is the *observation equation*:

$$Y_t = F_t^{\mathsf{T}} \theta_t + \nu_t, \quad \nu_t \sim N(0, V) \quad , \tag{4}$$

the other is the *evolution equation*:

$$\theta_t = G\theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, W) \quad . \tag{5}$$

Algorithm 1: Updating a DLM given G, V, W.

Initialize $t = 0$
$\{$Initial information $p(\theta_0|D_0) \sim \mathrm{N}[m_0, C_0]\}$
Input: m_0, C_0, G, V, W
loop
  $t = t + 1$
  $\{$Compute prior at t: $p(\theta_t|D_{t-1}) \sim \mathrm{N}[a_t, R_t]\}$
    $a_t = Gm_{t-1}$
    $R_t = GC_{t-1}G^{\mathsf{T}} + W$
  Input: F_t
  $\{$Compute forecast at t: $p(Y_t|D_{t-1}) \sim \mathrm{N}[f_t, Q_t]\}$
    $f_t = F_t^{\mathsf{T}} a_t$
    $Q_t = F_t^{\mathsf{T}} R_t F_t + V$
  Input: Y_t
  $\{$Compute forecast error $e_t\}$
    $e_t = Y_t - f_t$
  $\{$Compute adaptive vector $A_t\}$
    $A_t = R_t F_t Q_t^{-1}$
  $\{$Compute posterior at t: $p(\theta_t|D_t) \sim \mathrm{N}[m_t, C_t]\}$
    $m_t = a_t + A_t e_t$
    $C_t = R_t - A_t Q_t A_t^{\mathsf{T}}$
end loop

F_t^{T} is a row in the *design matrix* representing independent variables effecting Y_t. G is the *evolution matrix*, capturing deterministic changes to θ, where $\theta_t \approx G\theta_{t-1}$. V is the *observational variance*, $\mathrm{Var}(\varepsilon)$ in ordinary least squares. W is the *evolution variance matrix*, capturing random changes to θ, where $\theta_t = G\theta_{t-1} + w_t, \quad w_t \sim N(0, W)$. The two parameters G and W make a linear model *dynamic*.

2.2 Updating a DLM

The Bayesian nature of a DLM is evident in the careful accounting of sources of variation that generally increase system uncertainty; and, information in the form of incremental observations that generally decrease system uncertainty. A DLM starts with initial information, summarized by the parameters of a (frequently multivariate) normal distribution:

$$p(\theta_0|D_0) \sim N(m_0, C_0) \quad . \tag{6}$$

At each time step, the information is augmented as follows:

$$D_t = \{Y_t, D_{t-1}\} \quad . \tag{7}$$

Algorithm 1 details the relatively simple steps of updating a DLM as additional regression vectors F_t and observations Y_t become available. Note that upon arrival of the current regression vector F_t, a one-step forecast distribution $p(Y_t|D_{t-1})$ is computed using the prior distribution $p(\theta_t|D_{t-1})$, the regression vector F_t, and the observation noise V.

2.3 Model Likelihood

The one-step forecast distribution facilitates computation of *model likelihood* by evaluation of the density of the one-step forecast distribution $p(Y_t|D_{t-1})$ for observation Y_t. The distribution $p(Y_t|D_{t-1})$ is explicitly a function of the previous periods information D_{t-1}; and, implicitly a function of static model parameters $\{G,V,W\}$ and model state determined by a series of updates resulting from the history D_{t-1}. Defining a model at time t as $M_t = \{G,V,W,D_{t-1}\}$, and explicitly displaying the M_t dependency in the one-step forecast distribution, we see that the one-step forecast distribution is equivalent to model likelihood[1]:

$$p(Y_t|D_{t-1}) = p(Y_t, D_{t-1}|D_{t-1}, M_t) = p(D_t|M_t) \quad (8)$$

Model likelihood, $p(D_t|M_t)$, will be an important input to our mixture model discussed below.

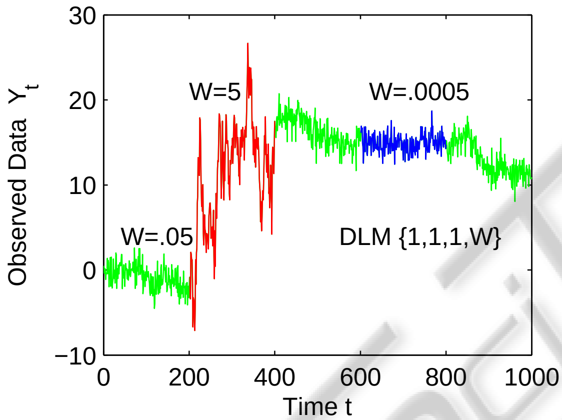

Figure 1: Observations Y_t generated from a mixture of three DLMs. Discussion appears in §2.4

2.4 Generating Observations

Before delving into mixtures of DLMs, we illustrate the effect of varying the evolution variance W on the state variable θ in a very simple DLM. In Figure 1 we define three very simple DLMs, $\{1,1,1,W_i\}, W_i \in \{.0005,.05,5\}$. The observations are from simple random walks, where the level of the series θ_t varies according to an evolution equation $\theta_t = \theta_{t-1} + \omega_t$, and the observation equation is $Y_t = \theta_t + \nu_t$. Compare the relative stability in the level of observations generated by the three models. Dramatic and interesting behavior materializes as W increases.

[1] $D_t = \{Y_t, D_{t-1}\}$ by definition; M_t contains D_{t-1} by definition; and, $p(Y_t, D_{t-1}|D_{t-1}) = p(Y_t|D_{t-1})p(D_{t-1}|D_{t-1}) = p(Y_t|D_{t-1})$.

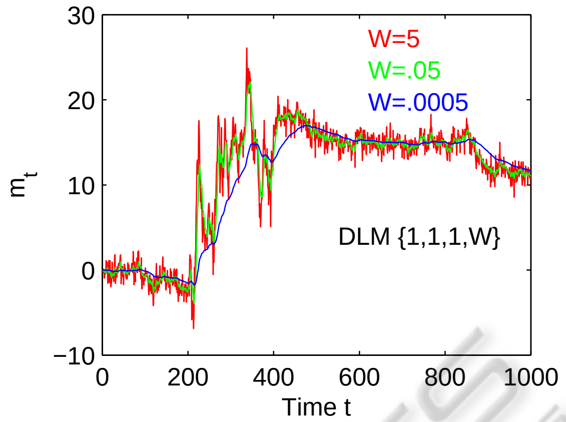

Figure 2: Estimates of the mean of the state variable θ_t for three DLMs when processing generated data of Figure 1.

2.5 Model Inference

Figure 1 illustrated the difference in appearance of observations Y_t generated with different DLM parameters. In Figure 2, note that models with smaller evolution variance W result in smoother estimates — at the expense of a delay in responding to changes in level. At the other end of the spectrum, large W permits rapid changes in estimates of θ — at the expense of smoothness. In terms of the model likelihood $p(D_t|M_t)$, if W is too small, the standardized forecast errors $e_t/\sqrt{Q_t}$ will be large in magnitude, and therefore model likelihood will be low. At the other extreme, if W is too large, the standardized forecast errors will appear small, but the model likelihood will be low now due to the diffuse forecast distribution.

In Figure 3, we graph the trailing interval log likelihoods for each of the three DLMs. We define trailing interval (k-period) likelihood as:

$$\begin{aligned} L_t(k) &= p(Y_t, Y_{t-1}, \ldots, Y_{t-k+1}|D_{t-k}) \\ &= p(Y_t|D_{t-1})p(Y_{t-1}|D_{t-2})\ldots \\ &\quad p(Y_{t-k+1}|D_{t-k}) \quad . \end{aligned} \quad (9)$$

This concept is very similar to Bayes' factors discussed in (West and Harrison, 1997), although we do not divide by the likelihood of an alternative model. Our trailing interval likelihood is also similar to the likelihood function discussed in (Crassidis and Cheng, 2007); but, we assume the errors e_t are not autocorrelated.

Across the top of Figure 3 appears a color code indicating the true model prevailing at time t. It is interesting to note when the likelihood of a model exceeds that of the true model. For instance, around the $t = 375$ mark, the model with the smallest evolution variance appears most likely. Reviewing Figure 2, the state estimates of DLM $\{1,1,1,W = .0005\}$ just

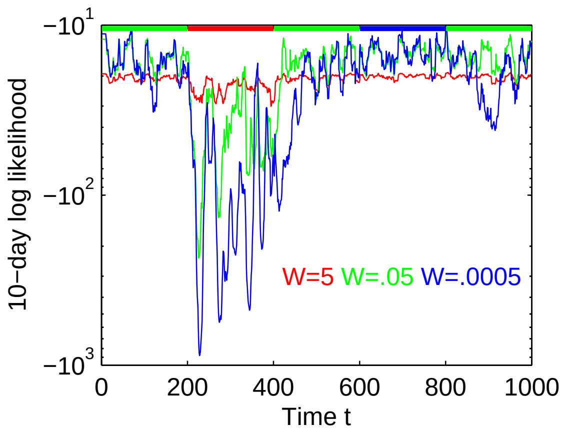

Figure 3: Log likelihood of observed data during most recent 10 days given the parameters of three DLMs when processing generated data of Figure 1. Bold band at top of figure indicates the true generating DLM.

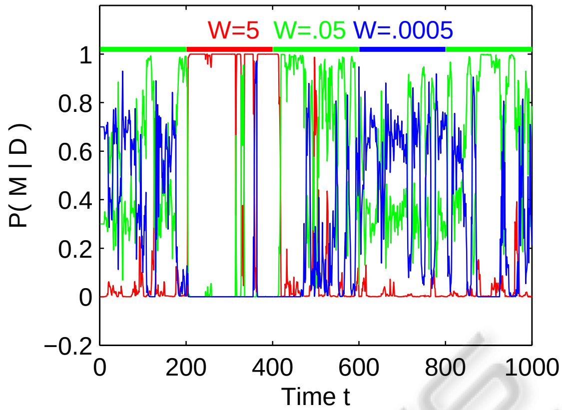

Figure 4: Model probabilities from normalized likelihoods of observed data during most recent 10 periods. Bold band at top of figure indicates the true generating DLM.

happened to be in the right place at the right time. Due to the more concentrated forecast distributions $p(Y_t|D_{t-1})$ of this model, it briefly attains the highest trailing 10-period log likelihood. A similar occurrence can be seen for the DLM $\{1,1,1,W=.05\}$ around $t=325$.

While the series on Figure 3 appear visually close at times, note the log scale. After converting back to normalized model probabilities, the favored model at a particular instance is more apparent as illustrated in Figure 4. In §5, we will perform model inference on the return series of exchange traded funds (ETFs).

3 PARAMETER ESTIMATION

In §2, we casually discussed DLMs varying in parameterization. Generating observations from a specified DLM or combination of DLMs, as in §2.4, is trivial. The inverse problem, determining model parameters from observations is significantly more challenging. There are two distinct versions of this task based upon area of application. In the simpler case, the parameters are unknown but assumed constant. A number of methods are available for model identification in this case, both off-line and on-line. For example, (Ghahramani and Hinton, 1996) use E-M off-line, and (Crassidis and Cheng, 2007) use the likelihood of a fixed-length trailing window of prediction errors on-line. Time varying parameters are significantly more challenging. The posterior distributions are path dependent and the number of paths is exponential in the length of the time series. Various approaches are invoked to obtain approximate solutions with reasonable computational effort. (West and Harrison, 1997) approximate the posterior with a single Gaussian that matches the moments of the exact distribution. (Valpola et al., 2004; Sarkka and Nummenmaa, 2009) propose variational Bayesian approximation. (Minka, T.P., 2007) discusses Gaussian-sum and assumed-density filters.

3.1 Multi-process Mixture Models

(West and Harrison, 1997) define sets of DLMs, where the defining parameters $M_t = \{F,G,V,W\}_t$ are indexed by λ^2, so that $M_t = M(\lambda_t)$. The set of DLMs at time t is $\{M(\lambda_t) : \lambda_t \in \Lambda\}$. Two types of multi-process models are defined. A *class I multi-process model*, where for some unknown $\lambda_0 \in \Lambda, M(\lambda_0)$ holds for all t; and, a *class II multi-process model* for some unknown sequence $\lambda_t \in \Lambda, (t=1,2,\ldots), M(\lambda_t)$ holds at time t. We build our model in §4 in the framework of a class II mixture model. We do not expect to be able to specify parameters exactly or finitely. Instead, we specify a set of models that quantize a range of values. In the terminology of (Sarkka and Nummenmaa, 2009), we will create a *grid approximation* to the evolution and observation variance distributions.

Class II mixture models permit the specification of a model per time period, leading to a number of potential model sequences exponential in the steps, $|\Lambda|^T$. However, in the spirit of the localized nature of dynamic models and practicality, (West and Harrison, 1997) exploit the fact that the value of information decreases quickly with time, and propose collapsing

[2](West and Harrison, 1997) index the set of component models $\alpha \in \mathcal{A}$; however, by convention in finance, α refers to stock specific return, consistent with §1.2. To avoid confusion, we index the set of component models $\lambda \in \Lambda$, consistent with the notation of (Chen and Liu, 2000).

the paths and approximating common posterior distributions. In the filtering literature, this technique is referred to as the *interacting multiple model (IMM) estimator* (Bar-Shalom et al., 2001, Ch. 11.6.6). In our application, in §5, we limit our sequences to two steps, and approximate common posterior distributions by collapsing individual paths based on the most recent two component models. To restate this briefly, we model two step sequences — the component model M_{t-1} just exited, and the component model M_t now occupied. Thus, we consider $|\Lambda|^2$ sequences. Reviewing Algorithm 1, the only information required from $t-1$ is captured in the collapsed approximate posterior distribution $p(\theta_{t-1}|D_{t-1}) \sim N(m_{t-1}, C_{t-1})$ for each component model $\lambda_{t-1} \in \Lambda$ considered.

3.2 Specifying Model Priors

One key input to mixture models are the model priors. We have tried several approaches to this task before finding a method suitable for our statistical arbitrage modeling task in §5. The goal of our entire modeling process is to design a set of model priors $p(M(\lambda_t))$ and model likelihoods $p(D|M(\lambda_t))$ that yield in combination insightful model posterior distributions $p(M(\lambda_t)|D)$, permitting the computation of quantities of interest by summing over the model space $\lambda_t \in \Lambda$ at time t:

$$p(X_t|D_t) \propto \sum_{\lambda_t \in \Lambda} p(X_t|M(\lambda_t))p(M(\lambda_t)|D_t) \tag{10}$$

In the context of modeling ETF returns discussed in §5, the vastly different scales for the contribution of W and V to Q left our model likelihoods unresponsive to values of W. This unresponsiveness was due to the fact that parameter values W and V are of similar scale; however, a typical $|F_t|$ for this model is approximately 0.01, and therefore the respective contributions to the forecast variance $Q = F^\intercal RF + V = F^\intercal(GCG^\intercal + \mathbf{W})F + \mathbf{V}$ are of vastly different scales, $1:10{,}000$. Specifically, density of the likelihood $p(Y_t|D_{t-1}) \sim N(f_t, Q_t)$ is practically constant for varying W after the scaling by 0.01^2. The only knob left for us to twist is that of the model priors.

DLMs with static parameters embed evidence of recent model relevance in their one-step forecast distributions. In contrast, mixture model component DLMs move forward in time from posterior distributions that mask model performance. The situation is similar to the game *best ball* in golf. After each player hits the ball, all players' balls are moved to a best position as a group. Analogously, when collapsing posterior distributions, sequences originating from different paths are approximated with a common posterior based upon end-point model. While some of us may appreciate obfuscation of our golf skills, the obfuscation of model performance is problematic. Due to the variance scaling issues of our application, the path collapsing, common posterior density approximating technique destroys the accumulation of evidence in one-step forecast distributions for specific DLM parameterizations $\lambda \in \Lambda$. In our current implementation, we retain local evidence of model effectiveness by running a parallel set of standalone (not mixed) DLMs. Thus, the total number of models maintained is $|\Lambda|^2 + |\Lambda|$, and the computational complexity remains asymptotically constant. In our mixture model, we define model priors proportional to trailing interval likelihoods from the standalone DLMs. This methodology locally preserves evidence for individual models as shown in Figure 3 and Figure 4.

The posterior distributions $p(\theta_t|D_t)_{M(\lambda)}$ emitted by identically parameterized standalone and component DLMs differ in general. A standalone constant parameter DLM computes the prior $p(\theta_t|D_{t-1})_{M(\lambda_t)}$ as outlined in Algorithm 1 using its own posterior $p(\theta_{t-1}|D_{t-1})_{M(\lambda_t=\lambda_{t-1})}$. In contrast, component DLMs compute prior distributions using a weighted posterior:

$$p(\theta_{t-1}|D_{t-1})_{M(\lambda_t)} = \sum_{\lambda_{t-1}} p(M(\lambda_{t-1})|M(\lambda_t))p(\theta_{t-1}|D_{t-1})_{M(\lambda_{t-1})} \quad . \tag{11}$$

4 A FINANCIAL EXAMPLE

(Montana et al., 2009) proposed a model for the returns of the S&P 500 Index based upon the largest principal component of the underlying stock returns. In the form $Y = F^\intercal\theta + \varepsilon$ used throughout this paper,

$$Y = r_{\text{s\&p}}, \quad F = r_{\text{pc1}}, \text{ and } \quad \theta = \beta_{\text{pc1}}. \tag{12}$$

The target and explanatory data in (Montana et al., 2009) spanned January 1997 to October 2005. We propose the use of two alternative price series that are very similar in nature; but, publicly available, widely disseminated, and tradeable. The proposed alternative to the S&P Index is the *SPDR S&P 500 ETF* (trading symbol SPY). SPY is an ETF designed to mimic the performance of the S&P 500 Index(PDR Services LLC, 2010). The proposed alternative to the largest principal component series is the *Rydex S&P Equal Weight ETF* (trading symbol RSP). RSP is an ETF designed to mimic the performance of the S&P Equal Weight Index (Rydex Distributors, LLC, 2010). While perhaps not as obvious a pairing as S&P Index /

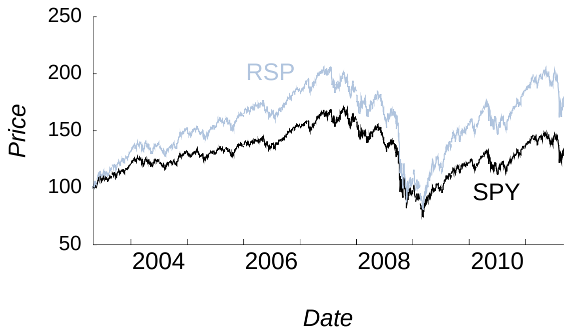

Figure 5: SPDR S&P 500 (SPY) and Rydex S&P Equal Weight (RSP) ETF closing prices, scaled to April 30, 2003 = 100.

SPY, a first principal component typically is the mean of the data — in our context, the mean is the equal weighted returns of the stocks underlying the S&P 500 Index. SPY began trading at the end of January 1993. RSP began trading at the end of April 2003. We use the daily closing prices P_t to compute daily log returns:

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right) \quad . \tag{13}$$

Our analysis is based on the months during which both ETFs traded, May 2003 to present (August 2011).

The price levels, scaled to 100 on April 30, 2003 are shown in Figure 5. Visually assessing the price series, it appears the two ETFs have common directions of movement, with RSP displaying somewhat greater range than SPY. Paralleling the work of (Montana et al., 2009), we will model the return of SPY as a linear function of RSP, $Y = F^{\mathsf{T}}\theta + \varepsilon$:

$$Y = r_{\text{spy}}, \quad F = r_{\text{rsp}}, \text{ and } \quad \theta = \beta_{\text{rsp}}. \tag{14}$$

We estimate the time varying regression parameter θ_t using a class II mixture model composed of 50 candidate models with parameters $\{F_t, 1, V, W\}$. $F_t = r_{\text{rsp}}$, the return of RSP, is common to all models. The observation variances are the values $V \times 1{,}000{,}000 \in \{\, 1,\ 2.15,\ 4.64,\ 10,\ 21.5,\ 46.4,\ 100,\ 215,\ 464,\ 1{,}000 \,\}$. The evolution variances are the values $W \times 1{,}000{,}000 \in \{\, 10,\ 56,\ 320,\ 1{,}800,\ 10{,}000 \,\}$. Our on-line process computes $50^2 + 50 = 2550$ DLMs, 50^2 DLMs corresponding to the two-period model sequences, and 50 standalone DLMs required for trailing interval likelihoods. In the mixture model, the priors $p(M(\lambda_t))$ for component models $M(\lambda_t)$, $\lambda_t \in \Lambda$, are proportional to trailing interval likelihoods (9) of corresponding identically parameterized standalone DLMs.

It's an interesting side topic to consider the potential scale of these mixtures. Circa 1989, in the

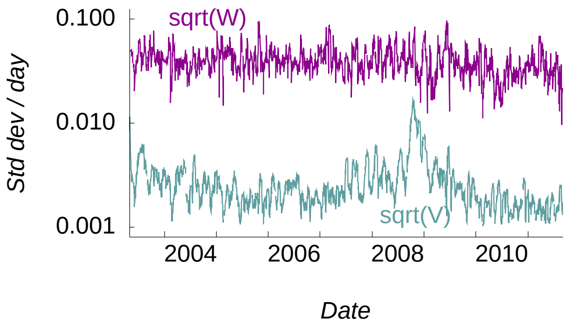

Figure 6: The daily standard deviation of ν_t and ω_t as estimated by the mixture model. Observation noise $\nu_t \sim N(0,V)$; evolution noise $\omega_t \sim N(0,W)$.

predecessor text to (West and Harrison, 1997), West and Harrison suggested the use of mixtures be restricted for purposes of "computational economy"; and that a single DLM would frequently be adequate. Approximately one decade later, (Yelland and Lee, 2003) were running a production forecasting system with 100 component models, and $10{,}000$ model sequence combinations. Now, more than two decades after West and Harrison's practical recommendation, with the advent of ubiquitous inexpensive GPGPUs, the economics of computation have changed dramatically. A direction of future research is to revisit implementation of large scale mixture models quantizing several dimensions simultaneously.

Subsequent to running the mixture model for the period May 2003 to present, we are able to review estimated time varying parameters V_t and W_t, as shown in Figure 6. This graph displays the standard deviation of observation and evolution noise, commonly referred to as volatility in the financial world. It is interesting to review the decomposition of this volatility. Whereas the relatively stationary series $\sqrt{W}$ in Figure 6 suggests the rate of evolution of θ_t is fairly constant across time; the observation variance V varies dramatically, rising noticeably during periods of financial stress in 2008 and 2009. The observation variance, or standard deviation as shown, may be interpreted as the end-of-day mispricing of SPY relative to RSP. In §5, we will demonstrate a trading strategy taking advantage of this mispricing. The increased observational variance at the end of 2008, visible in Figure 6 results in an increase in the rate of profitability of the statistical arbitrage application plainly visible in Figure 7.

5 STATISTICAL ARBITRAGE

(Montana et al., 2009) describe an illustrative statis-

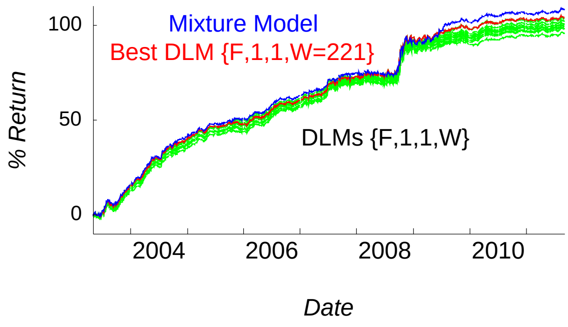

Figure 7: Cumulative return of the various implementations of a statistical arbitrage strategy based upon a time varying mixture model and 10 constant parameter DLMs.

tical arbitrage strategy. Their proposed strategy takes equal value trading positions opposite the sign of the most recently observed forecast error ε_{t-1}. In the terminology of this paper, they tested 11 constant parameter DLMs, with a parameterization variable δ equivalent to:

$$\delta = \frac{W}{W+V} \quad . \tag{15}$$

They note that this parameterization variable δ permits easy interpretation. With $\delta \approx 0$, results approach an ordinary least squares solution: $W = 0$ implies $\theta_t = \theta$. Alternatively, as δ moves from 0 towards 1, θ_t is increasingly permitted to vary.

Figure 6 challenges the concept that a constant specification of evolution and observation variance is appropriate for an ETF returns models. To explore the effectiveness of class II mixture models versus statically parameterized DLMs, we evaluated the performance of our mixture model against 10 constant parameter DLMs. We set $V = 1$ as did (Montana et al., 2009), and specified $W \in \{29, 61, 86, 109, 139, 179, 221, 280, 412, 739\}$. These values correspond to the 5, 15, ...95%-tile values of W/V observed in our mixture model.

Figure 6 offers no justification of using $V = 1$. While the prior $p(\theta_t|D_{t-1})$, one-step $p(Y_t|D_{t-1})$ and posterior $p(\theta_t|D_t)$ "distributions" emitted by these DLMs will not be meaningful, the intent of such a formulation is to provide time varying *point estimates* of the state vector θ_t. The *distribution* of θ_t is not of interest to modelers applying this approach. In the context of the statistical arbitrage application considered here, the distribution is not required. The trading rule proposed is based on the sign of the forecast error; and, the forecast is a function of the prior mean a_t (a point estimate) for the state vector θ_t and observed values F_t and Y_t: $\varepsilon_t = Y_t - F_t^{\top} a_t$.

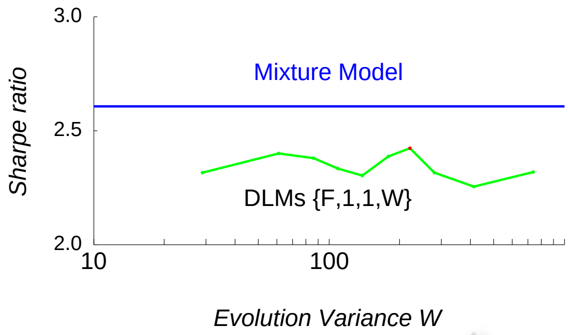

Figure 8: Sharpe ratios realized by the time varying mixture model and 10 constant parameter DLMs.

5.1 The Trading Strategy

Consistent with (Montana et al., 2009), we ignore trading and financing costs in this simplified experiment. Given the setup of constant absolute value SPY positions taken daily, we compute cumulative returns by summing the daily returns. The rule we implement is:

$$\texttt{portfolio}_t(\varepsilon_{t-1}) = \begin{cases} +1 & \text{if } \varepsilon_{t-1} \leq 0, \\ -1 & \text{if } \varepsilon_{t-1} > 0. \end{cases} \tag{16}$$

where $\texttt{portfolio}_t = +1$ denotes a long SPY and short RSP position; $\texttt{portfolio}_t = -1$ denotes a short SPY and long RSP position. The SPY leg of the trade is of constant magnitude. The RSP leg is $-a_t \times$ SPY-value, where a_t is the mean of the prior distribution of θ_t, $p(\theta_t|D_{t-1}) \sim N(a_t, R_t)$; and, recall from (14) the interpretation of θ_t is the sensitivity of the returns of SPY Y_t to the returns of RSP F_t. Note that this strategy is a modification to (Montana et al., 2009) in that we hedge the S&P exposure with the equal weighted ETF, attempting to capture mispricings while eliminating market exposure. The realized Sharpe ratios appear dramatically higher in all cases than in (Montana et al., 2009), primarily attributable to the hedging of market exposure in our variant of a simplified arbitrage example. Montana et al. report Sharpe ratios in the 0.4 - 0.8 range; in this paper, after inclusion of the hedging technique, Sharpe ratios are in the 2.3 - 2.6 range.

5.2 Analysis of Results

We reiterate that we did not include transaction costs in this simple example. Had we done so, the results would be significantly diminished. With that said, we will review the relative performance of the models for the trading application.

In Figure 7, it is striking that all models do fairly well. The strategy holds positions based upon a comparison of the returns of two ETFs, one scaled by an estimate of $\beta_{\mathtt{rsp},t}$. Apparently small variation in the estimates of the regression parameter are not of large consequence. Given the trading rule is based on the *sign* of the error ε_t, it appears that on many days, slight variation in the estimate of θ_t across DLMs does not result in a change to $\mathtt{sign}(\varepsilon_t)$. Figure 8 shows that over the interval studied, the mixture model provided a higher return per unit of risk, if only to a modest extent. What is worth mentioning is that the comparison we make is the *on-line* mixture model against the *ex post* best performance of all constant parameter models. Acknowledging this distinction, the mixture model's performance is more impressive.

6 CONCLUSIONS

Mixtures of dynamic linear models are a useful technology for modeling time series data. We show the ability of DLMs parameterized with time varying values to generate observations for complex dynamic processes. Using a mixture of DLMs, we extract time varying parameter estimates that offered insight to the returns process of the S&P 500 ETF during the financial crisis of 2008. Our *on-line* mixture model demonstrated superior performance compared to the *ex post* optimal component DLM in a statistical arbitrage application.

The contributions of this paper include the proposal of a method, trailing interval likelihood, for constructing component model prior probabilities. This technique facilitated successful modeling of time varying observational and evolution variance parameters, and captured model evidence not adequately conveyed in the one-step forecast distribution due to scaling issues. We proposed the use of two widely available time-series to facilitate easier replication and extension of the statistical arbitrage application proposed by (Montana et al., 2009). Our addition of a hedge to the statistical arbitrage application from (Montana et al., 2009) resulted in dramatically improved Sharpe ratios.

We have only scratched the surface of the modeling possibilities with DLMs. The mixture model technique eliminates the burden of *a priori* specification of process parameters. We look forward to evaluating models with higher dimension state vectors and parameterized evolution matrices. Due to the inherently parallel nature of DLM mixtures, we also look forward to exploring the ability of current hardware to tackle additional challenging modeling problems.

REFERENCES

Bar-Shalom, Y., Li, X., Kirubarajan, T., and Wiley, J. (2001). *Estimation with applications to tracking and navigation*. John Wiley & Sons, Inc.

Bishop, C. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer Science+Business Media, LLC. New York, NY, USA.

Chen, R. and Liu, J. (2000). Mixture Kalman filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):493–508.

Crassidis, J. and Cheng, Y. (2007). Generalized Multiple-Model Adaptive Estimation Using an Autocorrelation Approach. In *Information Fusion, 2006 9th International Conference on*, pages 1–8. IEEE.

Ghahramani, Z. and Hinton, G. (1996). Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, University of Toronto.

Hamilton, J. (1994). *Time series analysis*. Princeton University Press: Princeton, NJ, USA.

Johnson, R. and Wichern, D. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall: Upper Saddle River, NJ, USA.

Kalaba, R. and Tesfatsion, L. (1996). A multicriteria approach to model specification and estimation. *Computational Statistics & Data Analysis*, 21(2):193–214.

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

Minka, T. (1999). From hidden Markov models to linear dynamical systems. Technical Report 531, Vision and Modeling Group of Media Lab, MIT.

Minka, T.P. (2007). Bayesian inference in dynamic models: an overview. http://research.microsoft.com.

Montana, G., Triantafyllopoulos, K., and Tsagaris, T. (2009). Flexible least squares for temporal data mining and statistical arbitrage. *Expert Systems with Applications*, 36(2):2819–2830.

PDR Services LLC (2010). Prospectus. SPDR S&P 500 ETF. https://www.spdrs.com.

Rydex Distributors, LLC (2010). Prospectus. Rydex S&P Equal Weight ETF. http://www.rydex-sgi.com/.

Sarkka, S. and Nummenmaa, A. (2009). Recursive noise adaptive Kalman filtering by variational Bayesian approximations. *Automatic Control, IEEE Transactions on*, 54(3):596–600.

Valpola, H., Harva, M., and Karhunen, J. (2004). Hierarchical models of variance sources. *Signal Processing*, 84(2):267–282.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag New York, Inc. New York, NY, USA.

Yelland, P. and Lee, E. (2003). Forecasting product sales with dynamic linear mixture models. Technical Report SMLI TR-2003-122, Sun Microsystems, Inc.