

Forecasting Product Sales with Dynamic Linear Mixture Models

Phillip M. Yelland and Eunice Lee

Forecasting Product Sales with Dynamic Linear Mixture Models

Phillip M. Yelland
Eunice Lee

SMLI TR-2003-122

March 2003

Abstract:

This paper discusses investigations undertaken at Sun Microsystems, Inc. into practical forecasting applications of the Bayesian Dynamic Linear Models described in the seminal work of West and Harrison (1997). Particular emphasis is placed on the use of *class II mixture models*, which use Bayesian model averaging to help accommodate model uncertainty.

Keywords: Forecasting, probability, statistics, time-series.



M/S MTV29-01
2600 Casey Avenue
Mountain View, CA 94043

email address:
phillip.yelland@sun.com
eunice.lee@sun.com

© 2003 Sun Microsystems, Inc. All rights reserved. The SML Technical Report Series is published by Sun Microsystems Laboratories, of Sun Microsystems, Inc. Printed in U.S.A.

Unlimited copying without fee is permitted provided that the copies are not made nor distributed for direct commercial advantage, and credit to the source is given. Otherwise, no part of this work covered by copyright hereon may be reproduced in any form or by any means graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an information retrieval system, without the prior written permission of the copyright owner.

TRADEMARKS

Sun, Sun Microsystems, the Sun logo, and Java are trademarks or registered trademarks of Sun Microsystems, Inc. in the U.S. and other countries. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the U.S. and other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc. UNIX is a registered trademark in the United States and other countries, exclusively licensed through X/Open Company, Ltd.

For information regarding the SML Technical Report Series, contact Jeanie Treichel, Editor-in-Chief <jeanie.treichel@eng.sun.com>. All technical reports are available online on our Website, <http://research.sun.com/techrep/>.

Forecasting Product Sales With Dynamic Linear Mixture Models

Phillip M. Yelland
Sun Microsystems Laboratories
Mountain View, California

Eunice Lee
Sun Microsystems Inc. WorldWide Operations
Newark, California

Introduction

This paper describes some of our experiences at Sun Microsystems, Inc. applying a particular type of statistical model for time-series analysis to the forecasting of product sales. The challenges facing a prospective forecaster are illustrated with representative product sales histories. We indicate how the peculiar characteristics of such sales histories render traditional approaches to forecasting largely ineffectual.

Next, we introduce the *dynamic linear model (DLM)*, a Bayesian device for time-series analysis detailed in (Pole *et al.*, 1994) and (West and Harrison, 1997). The conceptual basis and practical details of the model are outlined, as are procedures for updating the model as new data comes to light. Unfortunately, it transpires that the performance of the DLM when applied to Sun's forecasting problems is as unimpressive as that of the traditional approaches.

An elaboration of the DLM, however, proves far more capable: a *class II mixture model* contains a collection of DLM's that it applies in combination to account for uncertainty in model specification and changes in regime. The efficacy of class II mixture models is demonstrated with respect to the sample series.

We conclude with a brief discussion of the implementation of a forecasting system based on class II mixture models at Sun Microsystems, and of directions for future work.

Patterns of Sales

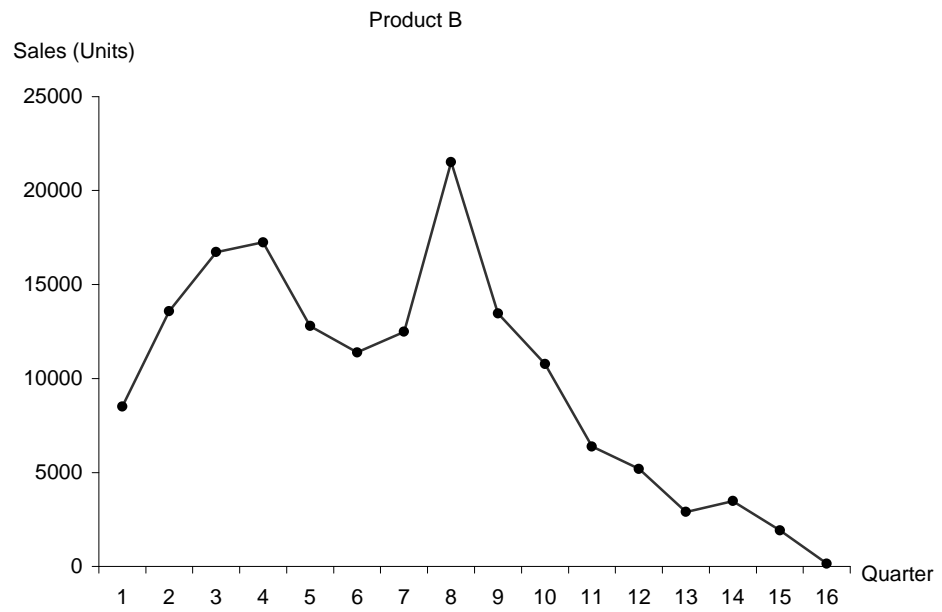
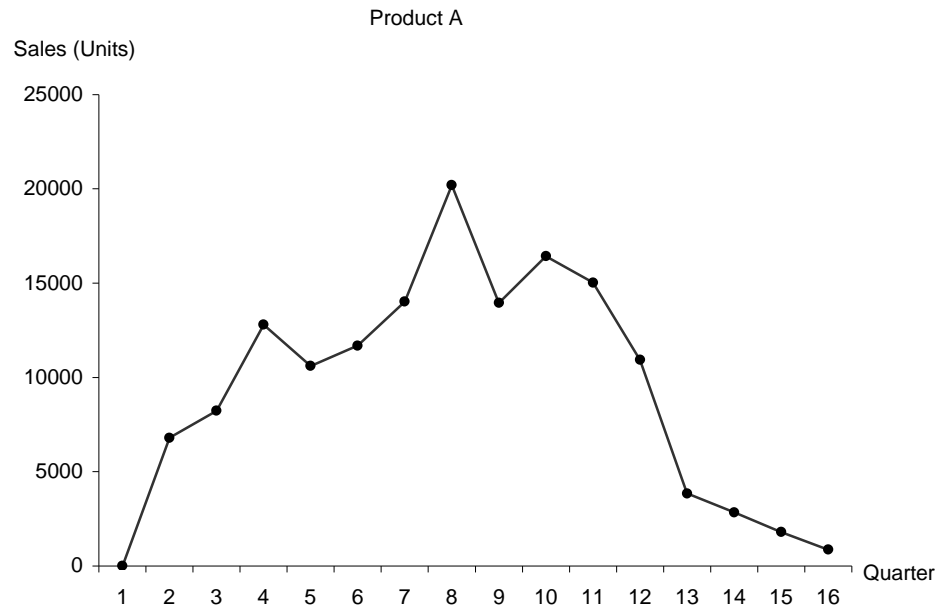


Figure 1: Sales for typical product lifecycles

Figure 1 illustrates the quarterly sales of two reasonably typical Sun products as they progress through their lifecycles.¹ Of course, not all products conform closely to these patterns of sales, but the series in the figure do exhibit characteristics displayed by many of the Company's products:

- Largely due to the rapid rate of product innovation in the technology sector, most lifecycles are fairly short. At around 16 quarters, these series are actually amongst the longest in the Company's product portfolio. Furthermore, organizational requirements demand forecasts as early as possible in the product lifecycle. A forecasting methodology, for example, that requires 10 quarters (i.e., 2½ years) or so of sales history before making its first reliable forecast is of very limited appeal.
- The series generally exhibit changes in regime or "structural breaks" (Clements and Hendry, 2001), manifest in periods during which the pattern of sales shifts profoundly. For instance, certain segments of the series exhibit seasonal variation in the form of a sales surge towards the end of the financial year, but this pattern is not consistent throughout the lifecycle.
- Though not apparent from the figure, sales are affected by a multitude of (possibly confounded) factors, many of them unknown at the time of forecasting. For example, it is clear in hindsight that periods 8 – 10 witnessed a profound diminution of overall sales in the markets for both products; it's not clear, however, if these market shifts pre-empted the normal drop-off in sales expected as a product nears the end of its lifecycle.

In combination, these characteristics frustrate attempts to apply many techniques for time-series analysis. Classical statistical approaches such as Box-Jenkins modeling (Makridakis *et al.*, 1997) are confuted by the series' brevity and unpredictable regime changes. Models of product lifecycles (Lilien *et al.* 1992) are undermined by the confounding effects of market shifts. Econometric regression models (Dinardo *et al.* 1996) are limited by a dearth of reliable leading indicators (indexes of expected GDP or sector-specific output, for example, proved to be broadly inapplicable, and even sales-force forecasts have failed to be consistently predictive).

We found only a few techniques of time-series analysis appropriate to the forecasting of the Sun's product sales: Simple smoothing approaches (which, as Diebold (2000) points out, share the virtue of functioning in many situations beyond the scope of more sophisticated methods), and — naturally enough — dynamic linear models. The next section looks at the application of the sim-

¹ While the examples in this paper are based on actual products, considerations of commercial confidentiality require that sales data be disguised.

ple smoothing techniques; while they turn out to be of only limited effectiveness in our context, they do comprise a useful baseline against which the DLM may be assessed.

Forecasting Using Smoothing Techniques

We experimented with three commonly used smoothing techniques:

- *Moving average*

This computes a prediction for the next value of the series that is simply the arithmetic mean of the last two observed values:

$$\hat{Y}_{t+1} = \frac{1}{2}(Y_t + Y_{t-1})$$

- *Exponential smoothing*

Produces a prediction as a weighted average (determined by the specified coefficient α) of the last observed value and the last prediction:

$$\hat{Y}_{t+1} = \alpha Y_t + (1 - \alpha) \hat{Y}_t$$

- *Holt-Winters with seasonality*

In this method, the prediction is formed by adding together two quantities representing the current level and trend of the series, and then multiplying by a proportion intended to capture the effects of (multiplicative) seasonality. Level, trend and seasonal effects (the latter with seasonal period p) are updated using separate exponential smoothing recurrences determined by coefficients α , β and γ :

$$\begin{aligned}\hat{Y}_{t+1} &= S_{t-p+1}(M_t + T_t), \\ M_t &= \alpha(M_{t-1} + T_{t-1}) + (1 - \alpha) \frac{Y_t}{S_{t-p}}, \\ T_t &= \beta T_{t-1} + (1 - \beta)(M_t - M_{t-1}), \\ S_t &= \gamma S_{t-p} + (1 - \gamma) \frac{Y_t}{M_t}.\end{aligned}$$

In all three cases, starting values for the recurrences are derived from the first few values of the series.

Forecast Metrics

“One-step ahead validation” (Makridakis *et al.*, 1997, Diebold, 2000) is an intuitively appealing means of appraising a forecasting method, given historical values for the time-series to which it is to be applied: For actual values Y_1, \dots, Y_n , for each value $i = i_0 + 1, \dots, n$ (where i_0 is a value large enough to allow the model to be calibrated at all), we calibrate the forecasting model using the

series Y_1, \dots, Y_{i-1} , and then compare the forecast value \hat{Y}_i with the actual value (or *observation*) Y_i . There is a plethora of measurements for comparing forecast and actual values, no one of which is uniformly superior to all the others. For the sake of brevity in presentation, we've chosen two:

- The *mean absolute deviation* (MAD) is simply the arithmetic mean of the absolute differences between predicted and actual values:²

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^n |\hat{Y}_i - Y_i|$$

- *Theil's U statistic* (Makridakis *et al.*, 1997) compares the performance of the forecasting method with that of the “naïve” forecast, which simply predicts that the next value in the series will equal the last observation. The comparison takes the form of a ratio of the corresponding root mean squared errors (the square root of the average squared differences between prediction and observation). As a rule of thumb, a forecasting method that yields a Theil's U of greater than one is judged ineffective.

$$U = \sqrt{\frac{\sum_1^T (\hat{Y}_t - Y_t)^2}{\sum_1^T (Y_t - Y_{t-1})^2}}$$

Table 1 summarizes the performance of the smoothing techniques described above for the three products introduced in the previous section. In each case, an attempt was made to select the smoothing parameters that presented each technique in the best light. Overall, the application of these simple forecasting techniques to Sun's quarterly sales data yields less than satisfactory results — none of the forecasts has a Theil's U less than 1 (the minimum for effectuality).

Technique	Product			
	A		B	
	U	MAD	U	MAD
Moving average	1.11	3159.1	1.31	3550.9
Exponential smoothing ³	1.00	3129.5	1.00	3097.9
Holt-Winters ⁴	1.52	4408.1	1.39	3826.9

Table 1: Performance of smoothing techniques

² Since Sun's inventory risks are roughly linear in the number of units held, one might reasonably contend that the use of the MAD is better motivated than — for example — the root mean square error.

³ Model parameters: $\alpha = 0.9$.

⁴ Model parameters: $\alpha = 0.3$, $\beta = 0.9$, $\delta = 0$.

The Dynamic Linear Model

The Dynamic Linear Model is a development of the so-called “state-space” approach to the estimation and control of dynamic systems (Aplevich, 1999). A state-space model of a time-series comprises a data generating process with a *state* (normally expressed as a vector of parameters) that may change over time. This state is only indirectly observed, in as far as the values of the time-series are derived as a function of the state in the corresponding period. Forecasting using a state-space model involves reconstructing the progress of the state from historical data, and then deriving future values of the time-series by extrapolating the state’s trajectory. Many state-space techniques (the DLM among them) derive from the *Kalman Filter*, which has found applications in fields as diverse as econometrics and inertial navigation. A thorough treatment of the Kalman Filter may be found in (Grewal and Andrews, 2001), and a comprehensive discussion of state-space models in forecasting is given in (Harvey, 1994).

While many state-space methods (such as those in (Harvey, 1994)) rely on classical statistical techniques, such as maximum likelihood estimation, the Dynamic Linear Model — though sharing many of the same underpinnings — is based on Bayesian statistical reasoning (Bernardo and Smith, 1994). The most complete exposition of the DLM is (West and Harrison, 1997) — much of the technical detail touched on in this paper is expanded upon in that work. A briefer introduction to the area, with a greater emphasis on practical applications, is (Pole *et al.*, 1994).

Predictive performance aside, state-space models have several features that recommend them for our application:

- The state-space formulation is particularly general; it is not difficult to devise state-space analogs of moving average, exponential smoothing, Holt-Winters or even ARIMA models (Makridakis *et al.*, 1997). Such flexibility permits the exploration of a wide variety of approaches using the same underlying framework.
- In many models, states have readily interpretable components — in those used in this paper, an underlying level of demand, trend in demand and seasonal perturbations may be distinguished. This facilitates communication with the users of the forecast models.

With its Bayesian underpinnings, the DLM in particular has further advantages:

- With suitably informative priors (Bernardo and Smith, 1994), forecasts can be produced for recently introduced products with insufficient sales histories to allow for purely data-driven calibration.

- The DLM is “open” in the sense of (Clements and Hendry, 2001): (West and Harrison, 1997) and (Pole *et al.*, 1994) show how the DLM allows additional extrinsic information to be incorporated into the model as it becomes available. By this means, for example, the introduction of a competing product (by another company or by Sun itself) may be anticipated.
- By making the structure of the model itself an uncertain quantity or “random variable” (a maneuver cheerfully sanctioned by Bayesian practice), changes in regime are propitiously addressed (as described in greater detail below).

The Model in Detail

Central to the DLM is a data generating process that evolves over time. At time t , the state of this process is expressed as a vector of parameters, $\boldsymbol{\theta}_t$. The corresponding value of the time-series — the observation Y_t — is a linear function of these parameters specified by the *regression vector* \mathbf{F}_t and a random noise term v_t . The latter is assumed to be simply normally distributed noise with known variance.⁵

The course of the process evolution is determined by the *evolution matrix* G_t , which pre-multiplies the previous period’s state vector. An *evolution noise* vector $\boldsymbol{\omega}_t$ is added to the result to produce the state vector in the current period. The noise vector has a multivariate normal distribution, and is serially uncorrelated and uncorrelated with the observation noise. Starting values for the state are given by a multivariate normal distribution.

This is summarized below; the notation $x \sim \mathbf{N}[\boldsymbol{\mu}, \boldsymbol{\sigma}^2]$ indicates a normally distributed random variable with the given mean and variance (the notation generalizes straightforwardly to random vectors with multivariate normal distributions), and \mathbf{F}_t' is the transpose of vector \mathbf{F}_t :

$$\begin{array}{lll}
 \text{Observation equation:} & Y_t = \mathbf{F}_t' \boldsymbol{\theta}_t + v_t, & v_t \sim \mathbf{N}[0, V_t], \\
 \text{System equation:} & \boldsymbol{\theta}_t = \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, & \boldsymbol{\omega}_t \sim \mathbf{N}[\mathbf{0}, \mathbf{W}_t], \\
 \text{Initial information:} & \boldsymbol{\theta}_0 \sim \mathbf{N}[\mathbf{m}_0, \mathbf{C}_0] &
 \end{array}$$

Gathering the parameters of such a model together in a tuple constitutes the DLM specification $\langle \mathbf{F}_t, \mathbf{G}_t, V_t, \mathbf{W}_t \rangle$.

⁵ In fact, the system used to make the forecasts in this paper actually makes provision for an unknown observation noise variance, but assuming a known variance considerably simplifies the presentation. Details of so-called *variance learning* may be found in (West and Harrison, 1997) or (Pole *et al.*, 1994).

Component Forms for DLM's

In order to specify a DLM, parameters \mathbf{F}_t , \mathbf{G}_t , \mathbf{V}_t , and \mathbf{W}_t must be specified for each period t . As stated earlier, this paper glosses over the specification of \mathbf{V}_t — full details are given in (West and Harrison, 1997) and (Pole *et al.*, 1994). Specification of \mathbf{W}_t is accomplished using *discount factors*, as outlined in the next section. To see how \mathbf{F}_t and \mathbf{G}_t may be formulated, we consider a number of elementary DLM's, each intended to capture a single aspect of a time-series.

Polynomial models begin with the simple *local level* model.⁶ Here, the process state comprises a single random variable (the expected value of the corresponding observation) engaged in a random walk over time. In symbols:

$$\begin{aligned} Y_t &= \theta_t + v_t, & v_t &\sim N[0, \mathbf{V}_t] \\ \theta_t &= \theta_{t-1} + \omega_t, & \omega_t &\sim N[0, \mathbf{W}_t] \end{aligned}$$

In the terms of the defining equations of the DLM, this implies a constant regression vector and evolution matrix with a single element:

$$\mathbf{F}_t = (1), \quad \mathbf{G}_t = (1)$$

The *first-order* polynomial or *local linear trend* model adds a growth component to the local level. The growth component itself drifts over time:

$$\begin{aligned} Y_t &= \theta_{1,t} + v_t, & v_t &\sim N[0, \mathbf{V}_t] \\ \theta_{1,t} &= \theta_{1,t-1} + \theta_{2,t-1} + \omega_{1,t}, & \omega_{1,t} &\sim N[0, \mathbf{W}_t] \\ \theta_{2,t} &= \theta_{2,t-1} + \omega_{2,t} \end{aligned}$$

Again, this gives rise to corresponding matrices:

$$\mathbf{F}_t = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{G}_t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

A straightforward extension of the formulations above produces higher-order polynomial models (West and Harrison, 1997).

Seasonality may be expressed in a number of ways using a DLM — again, see (West and Harrison, 1997) or (Pole *et al.*, 1994) for details. Seeking greatest flexibility, we elected to use a system of *seasonal dummies* (Diebold, 2000). To illustrate this approach, imagine a process with a default expected level of zero, with a seasonal cycle of n periods, and with seasonal variations in periods $\{s_1, \dots, s_m\} \subseteq \{1, \dots, n\}$. (An example of such a process might be one with a twelve-month

⁶ This model is also known as the *simple random walk with noise* (Enders, 1995).

cycle and seasonal variation in months $s_1 = 5$ and $s_2 = 10$.) The corresponding DLM has a constant evolution matrix:

$$\mathbf{G}_t = \mathbf{I}_d(n), \text{ the } n \times n \text{ identity matrix}$$

The regression vector \mathbf{F}_t changes from period to period according to the seasonal structure:

$$\mathbf{F}_t = \begin{cases} \mathbf{F}(i, n) & \text{if } ((t-1) \bmod n) + 1 = s_i \\ \mathbf{F}(0, n) & \text{otherwise} \end{cases}$$

Here:

$$\mathbf{F}(i, n) = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}, \text{ where } f_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

More complex DLM's may be composed by the *superposition* of components of the above forms. Consider DLM's $1, \dots, m$, so that at time t , DLM i has regression vector and evolution matrix $\mathbf{F}_{i,t}$ and $\mathbf{G}_{i,t}$ respectively. Superposing these DLM's produces a single DLM such that:

$$\mathbf{F}_t = \begin{pmatrix} \mathbf{F}_{1,t} \\ \vdots \\ \mathbf{F}_{m,t} \end{pmatrix} \text{ and } \mathbf{G}_t = \begin{pmatrix} \mathbf{G}_{1,t} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{G}_{m,t} \end{pmatrix}$$

Using superposition, for example, we are able to take a local linear trend model and a seasonal model and produce a DLM expressing a local trend *and* seasonality.

Estimation and Forecasting

Given a specified DLM, forecasting requires an estimate of the process state at the current time period. Given such an estimate, the evolution and observation equations of the DLM yield an estimate of the next observation — the forecast prediction. When and if the next observation is actually made available, a new, updated estimate of the process state may be derived.

Following the precepts of Bayesian reasoning (Bernardo and Smith, 1994), these estimates are expressed as probability distributions. Such distributions are conditional on a “current information set”, representing the sum of the knowledge about the process under observation at a particular time. In the usual instance, this information set consists of the initial state estimate (which you will recall takes the form of a multivariate normal distribution with specified mean and covariance) and the observations up to that period:

$$D_t = \{Y_t, \dots, Y_1, \mathbf{m}_0, \mathbf{C}_0\}$$

Then the process of forecasting revolves around the estimation of the three conditional distributions:

1. The *prior distribution* of θ_t , denoted $\theta_t | D_{t-1}$
This represents the model's "best guess" at the system's state after observing Y_1, \dots, Y_{t-1} , but before observing Y_t .
2. The *forecast distribution*, $Y_t | D_{t-1}$
The prediction for Y_t after observing Y_1, \dots, Y_{t-1} . As suggested above, this is derived from the prior distribution $\theta_t | D_{t-1}$.
3. The *posterior distribution* of θ_t , $\theta_t | D_t$
This results from the revision (using Bayes' theorem) of the prior distribution in the light of the observation of Y_t .

Technical details of the derivation of the above are given in (West and Harrison, 1997). For the purposes of the remainder of this paper, it is sufficient to note that thanks to the assumptions embodied in the DLM, all the required distributions are normal (univariate in the case of the forecast distribution, and multivariate for the state estimates).

Specification of Evolution Variance

In the exposition of the DLM above, the evolution noise covariance, \mathbf{W}_t , was left unspecified. Standard practice in the application of DLM's is the use of *discount factors* to determine the noise variance evolution indirectly. In a DLM with (a single) discount factor $\delta \in (0,1]$, the evolution noise covariance at time t is a fraction of the covariance of the posterior distribution of the state at time $t-1$:

$$\theta_{t-1} | D_{t-1} \sim \mathbf{N}[\mathbf{m}, \mathbf{C}] \quad \Rightarrow \quad \mathbf{W}_t = \frac{1-\delta}{\delta} \mathbf{C}$$

The discount factor regulates the rate of drift of the state vector in a reasonably intuitive fashion; a lower discount factor specifies a larger noise variance, allowing for a greater period-to-period change in the state vector.

When DLM's are constituted by the superposition of components, different discount factors are normally associated with each component. Thus a DLM with a level and a seasonal component, for example, would have one discount factor for the level, another for the seasonal variations. In such cases, the evolution noise covariance is a block diagonal matrix with a block for each component (computed from the part of the posterior covariance corresponding to that component).

Initial State Estimate

The final task in the specification of a dynamic linear model is the provision of an initial estimate for the process state (in the form of a multivariate normal distribution with given mean and covariance). In our use of the DLM, we approach this requirement in one of two ways:

- Where no sales history is available, estimates may be provided directly. Here, we make use of the results of formal decision analysis procedures used at Sun to produce estimates of initial product sales (Lee, 2002).
- If the product has sufficient sales history, we use the *reference analysis* of (Pole and West, 1989) to produce an estimate of the process state, dispensing with the need to provide one explicitly.

Performance of the DLM

In order to compare the performance of the DLM with the smoothing techniques examined earlier (and more importantly, to conform with Sun's current forecasting policies), a single point must be chosen from the forecast distribution produced by the model in each period. A sedulously Bayesian approach would require the definition of an appropriate utility function (Bernardo and Smith, 1994); unfortunately, fully characterizing the Company's manufacturing and sales processes with a view to defining such a function is an epic undertaking in itself. So we elected to use the median of the distribution as a point forecast. This corresponds to a utility function linear in the forecast error, which is not a bad approximation to inventory costs arising from overestimates (though not necessarily to the opportunity costs of lost sales from underestimates).

Table 2 compares the results of running a DLM⁷ on the two sample series with the results of the simple forecasting techniques shown earlier. Again, note that in both cases, the DLM has a Theil's U greater than 1.⁸

⁷ The particular model used was a local trend model with a discount factor of 0.9 and no seasonality; again, some experimentation was allowed to improve forecast performance.

⁸ Also note that there is no necessary contradiction in the fact that the MAD of the DLM is lower than that of exponential smoothing, while its Theil's U statistic is higher.

Technique	Product			
	A		B	
	U	MAD	U	MAD
Moving average	1.11	3159.1	1.31	3550.9
Exponential smoothing	1.00	3129.5	1.00	3097.9
Holt-Winters	1.52	4408.1	1.39	3826.9
<i>DLM</i>	<i>1.04</i>	<i>2873.9</i>	<i>1.07</i>	<i>2999</i>

Table 2: Performance of the DLM

Mixture Processes

The performance of the basic DLM displayed in the previous section is disappointing, particularly in light of its relative complexity compared with the smoothing techniques described earlier. Fortunately, it is possible to improve significantly on the basic DLM's performance by incorporating it into a broader framework.

As West and Harrison (1997) point out, there are two chief sources of forecast inaccuracy in the basic DLM:

1. It is normally unreasonable to suppose that a single DLM should represent a time-series throughout its entire duration. This is particularly the case in the series such as those dealt with here, which are subject to marked changes in regime.
2. The specification of a single DLM fails to represent the degree of *model uncertainty* (Draper 1995) associated with the specification. In actuality, of course, a range of specifications may apply to a series — even in a single period. While it might be expected that properly accounting for such uncertainty would degrade the predictive performance of a forecasting technique (since forecasts would be “more tentative”), Draper (1997) shows that in fact, accurate treatment of model uncertainty can *increase* forecast performance.

The first of these shortcomings may be addressed in an ad-hoc manner, using what West and Harrison term *automatic intervention*. This allows a single DLM partially to mimic a multi-regime model by examining the performance⁹ of the DLM retrospectively after each observation, switch-

⁹ Performance may be characterized using a variety of techniques; see the hypothesis testing approaches in (Gelman *et al.* 1995), for example. West and Harrison (1997) suggest a “cusum” approach that recapitulates the “sequential probability ratio tests” of Wald (1947).

ing to an “intervention regime” (normally with radically different model parameters, but with the same model components) in the event performance drops below a given threshold.

(West and Harrison, 1997) describes a more comprehensive solution, however, which promises to allay both problems; it is introduced below.

Class II Mixture Processes

Informally speaking, a *class II mixture process (MP)* is a composite model containing a collection of DLM’s. Each DLM is intended to characterize a regime in the series under scrutiny, applicable with a given (fixed) probability in each period. For example, one component DLM (intended to characterize periods of relative stability in a series) might have a discount factor of 0.9 (thus limiting the period-to-period drift), while another (addressing periods of greater volatility) has a discount factor of 0.4. The state of an MP expresses belief not only concerning the parameters of its component DLM’s, but the belief as to which component DLM applies in a particular period.

In general, suppose that we are given an MP model with k component DLM’s. Inference in the MP model is based on assumptions as to which particular component model applies to each period under consideration. Let the index $j_\tau \in \{1, \dots, k\}$ designate the component DLM that it is supposed applies at time τ . Then inferences are based upon assumptions relating to particular sequences of models $j_t, j_{t-1}, \dots, j_2, j_1$, implying that model j_1 applies at time 1, j_2 at time 2, and so on.

We can derive *conditional* prior, forecast and posterior distributions based on the assumption of a particular sequence of models using the DLM updating procedures outlined above. Consider, for example, deriving the conditional forecast distribution:

$$p(Y_t | j_t, \dots, j_1, D_{t-1})$$

This implies using model j_1 to estimate the (posterior) state at time 1, then using model j_2 to estimate the state at time 2, and so on, finally using model j_t to form the forecast distribution for Y_t .

To be useful in forecasting, however, we need to derive *unconditional* distributions, free of assumptions concerning model-sequences. In the usual Bayesian fashion, this is achieved by marginalizing model assumptions from a joint distribution (Bernardo and Smith, 1994). Since there are only k (disjoint) candidate models in each period, the rule of total probability gives:

$$p(Y_t | D_{t-1}) = \sum_{j_t=1}^k \sum_{j_{t-1}=1}^k \dots \sum_{j_1=1}^k p(Y_t, j_t, j_{t-1}, \dots, j_1 | D_{t-1})$$

The summand may be expressed in factors:

$$p(Y_t, j_t, j_{t-1}, \dots, j_1 | D_{t-1}) = p(Y_t | j_t, j_{t-1}, \dots, j_1, D_{t-1}) \times p(j_t, j_{t-1}, \dots, j_1 | D_{t-1})$$

The first term in the product above is simply the conditional forecast distribution discussed above. The second term is the probability of the model sequence j_t, j_{t-1}, \dots, j_1 estimated *before* observation of Y_t — or more succinctly, the *prior probability* of model sequence j_t, j_{t-1}, \dots, j_1 . It may be computed:

$$p(j_t, j_{t-1}, \dots, j_1 | D_{t-1}) = p(j_t | j_{t-1}, \dots, j_1, D_{t-1}) \times p(j_{t-1}, \dots, j_1 | D_{t-1})$$

In a class-II mixture process model, the conditional probability of j_t (regardless of the models that precede it) is simply a fixed constant, π_{j_t} , provided in the specification of the mixture model, so:

$$p(j_t, j_{t-1}, \dots, j_1 | D_{t-1}) = \pi_{j_t} p(j_{t-1}, \dots, j_1 | D_{t-1})$$

Finally, the *posterior* probability of model sequence j_{t-1}, \dots, j_1 (the latter term in the above) may be computed recursively by Bayes' rule:

$$\begin{aligned} p(j_{t-1}, j_{t-2}, \dots, j_1 | D_{t-1}) &= p(j_{t-1}, j_{t-2}, \dots, j_1 | Y_{t-1}, D_{t-2}) \\ &\propto p(Y_{t-1}, j_{t-1}, j_{t-2}, \dots, j_1 | D_{t-2}) \\ &= p(Y_{t-1} | j_{t-1}, j_{t-2}, \dots, j_1, D_{t-2}) p(j_{t-1} | j_{t-2}, \dots, j_1, D_{t-2}) p(j_{t-2}, \dots, j_1 | D_{t-2}) \\ &= p(Y_{t-1} | j_{t-1}, j_{t-2}, \dots, j_1, D_{t-2}) \pi_{j_{t-1}} p(j_{t-2}, \dots, j_1 | D_{t-2}) \end{aligned}$$

Note that the terms of the last right-hand side in the above are respectively the probability of Y_{t-1} according to the conditional forecast distribution, the fixed prior probability of model j_{t-1} , and the posterior probability of model sequence j_{t-2}, \dots, j_1 .

Approximation of Model Sequences

The analysis in the previous section requires exhaustive examination of all possible model sequences for all periods; clearly, for a non-trivial number of component models, the size of such a set quickly becomes intractable. Typically, for example, we use around 100 component DLM's (see below for more detail), and this approach would require consideration of about 10^{16} sequences after just 10 periods — the number of model sequences increases by a factor of 100 in every succeeding period.

As West and Harrison (1997) observe, however (and as West (1992) amplifies), it is possible to approximate the above analysis without significant loss of accuracy. More explicitly, there is a

fairly small lag h (usually less than 3) beyond which model selections normally have little effect on current inferences. So, for example, in the case of forecast distributions, we have:

$$p(Y_t | j_t, j_{t-1}, j_{t-h}, j_{t-h-1}, \dots, j_1, D_{t-1}) \approx p(Y_t | j_t, j_{t-1}, j_{t-h}, j'_{t-h-1}, \dots, j'_1, D_{t-1}),$$

even if $(j_{t-h-1}, \dots, j_1) \neq (j'_{t-h-1}, \dots, j'_1)$.

Similar results apply to prior and posterior distributions, too. So for the purposes of forecasting, distributions resulting from model sequences of the form $j_t, \dots, j_{t-h}, j_{t-h-1}, \dots, j_1$ (for all possible values of j_{t-h-1}, \dots, j_1) may be approximated by a single distribution¹⁰ associated with the prefix sequence j_t, \dots, j_{t-h} . The number of prefix sequences is much smaller than the number of complete sequences (just 10,000, for example, with 100 component models and $h = 2$), and does not grow exponentially with time.

Specification of Component Models

Of course, effective use of a mixture model like those described above requires propitious specification of the set of component models. As Draper (1995) indicates, such a set should contain a selection of models sufficient to describe the possible properties of the process in every conceivable regime.

We have found that a reasonably serviceable means of achieving this is simply to vary each particular aspect of the component specification independently, generating the component models from all possible combinations. For example, if a forecaster using a local level model with seasonality determines that the level discount δ_{level} may take values 0.2 or 0.9 (characterizing periods of relatively high and low variability respectively) while the seasonality discount δ_{season} is either 0.6 or 0.95 (patterns of seasonality normally being less volatile than those of levels), then a mixture model comprising the following components is generated:

$$\begin{aligned} &(\delta_{\text{level}} = 0.2, \delta_{\text{season}} = 0.6), \\ &(\delta_{\text{level}} = 0.9, \delta_{\text{season}} = 0.6), \\ &(\delta_{\text{level}} = 0.2, \delta_{\text{season}} = 0.95), \\ &(\delta_{\text{level}} = 0.9, \delta_{\text{season}} = 0.95). \end{aligned}$$

A simple elaboration of this technique also produces the model probabilities π_i used to compute the prior and posterior model sequence probabilities above: We may associate weights with each

¹⁰ The approximating distribution is determined by moments computed as averages of those associated with the sequences approximated, weighted by the posterior probabilities of those sequences; again, details may be found in (West and Harrison, 1997).

value of an attribute, reflecting the relative probabilities of each value in any period. Assuming independence of attributes, the overall probability of a given combination is simply the (normalized) product of the weights of its values. So in the example, if:

$$\begin{aligned} \text{weight}[\delta_{\text{level}} = 0.2] &= 1, & \text{weight}[\delta_{\text{level}} = 0.9] &= 3, \\ \text{weight}[\delta_{\text{season}} = 0.6] &= 2, & \text{weight}[\delta_{\text{season}} = 0.95] &= 3, \end{aligned}$$

then (noting that $2 + 6 + 3 + 9 = 20$):

$$\begin{aligned} \text{weight}[(\delta_{\text{level}} = 0.2, \delta_{\text{season}} = 0.6)] &= 1 \times 2 = 2 \Rightarrow \pi_1 = \frac{2}{20} = 0.1, \\ \text{weight}[(\delta_{\text{level}} = 0.9, \delta_{\text{season}} = 0.6)] &= 3 \times 2 = 6 \Rightarrow \pi_2 = \frac{6}{20} = 0.3, \\ \text{weight}[(\delta_{\text{level}} = 0.2, \delta_{\text{season}} = 0.95)] &= 1 \times 3 = 3 \Rightarrow \pi_3 = \frac{3}{20} = 0.15, \\ \text{weight}[(\delta_{\text{level}} = 0.9, \delta_{\text{season}} = 0.95)] &= 3 \times 3 = 9 \Rightarrow \pi_4 = \frac{9}{20} = 0.45. \end{aligned}$$

Performance of the Mixture Process Model

Table 3 compares the performance of earlier forecasting techniques with a single mixture process model comprising some 100 components.¹¹ The comparison is favorable; for both products, the Theil's U and Mean Absolute Deviation of the mixture model are less than those of any other technique. Furthermore, the Theil's U of the mixture model is less than 1 in both cases, indicating that the model is preferred to the naïve technique.

It should be observed that the computational demands of the mixture model may be significant; the results in Table 3 required approximately 30 minutes of processing time using a machine equipped with Sun's UltraSPARC[®] II CPU running at 400MHz with 512Mbytes of main memory.¹² Given that Sun's forecasts are produced on a two-weekly cycle, such considerations do not constitute an insurmountable obstacle in our circumstances; nonetheless, the technique can hardly be recommended in real-time or (more plausibly) interactive forecasting applications — at least in our current implementation.

¹¹ Attributes such as level, trend and seasonality discounts, as well as model order and seasonality structure varied across the components in the example. An advantage of the mixture process approach is that the same model specification may be used across a wide range of products. Thus in contrast to the techniques described earlier, no manual "model selection" procedure was applied to the mixture model to suit it to the sample series.

¹² Much of the computation required applies not to the manipulation of the component models themselves but rather to the complex mixture distributions that result from marginalizing model sequences during forecasting.

Technique	Product			
	A		B	
	U	MAD	U	MAD
Moving average	1.11	3159.1	1.31	3550.9
Exponential smoothing	1.00	3129.5	1.00	3097.9
Holt-Winters	1.52	4408.1	1.39	3826.9
DLM	1.04	2873.9	1.07	2999
<i>Mixture process model</i>	<i>0.66</i>	<i>1861.6</i>	<i>0.91</i>	<i>2937.4</i>

Table 3: Performance of the mixture process model

Conclusions and Future Work

We have shown how a class-II mixture process model is particularly well suited to forecasting sales demand for two products that are reasonably representative of many in Sun Microsystems' product range. In fact, the results of the previous section are quite representative of our experience with this technique — a suitably specified mixture model provides forecasts consistently superior to those of any other approach we've attempted.

The bulk of our current implementation is written using the Java™ platform (Gosling *et al.*, 2000) — largely in the interests of performance — with a front-end in Mathematica® (Wolfram, 1996). Extensive libraries written in the latter support the manipulation and display of time-series and forecasts. Such an arrangement makes effective use of our distributed computing environment (the main computation may be hosted by a different machine from that running the front-end), and supports the experimentation required in our initial applications. However, even routine use of the system requires the mediation of someone well versed in the details of the technique and its implementation, and we are in the process of completing a new user-friendly front-end (running in a Web-browser) that will support more widespread deployment.

References

- Aplevich, J., 1999. *The Essentials of Linear State-Space Systems*. J. Wiley and Sons.
- Bernardo, J., Smith, A., 1994. *Bayesian Theory*. J. Wiley and Sons.
- Clements, M., Hendry, D., 2001. *Forecasting Non-Stationary Economic Time-series*. MIT Press.
- Diebold, F., 2000. *Elements of Forecasting (2nd ed.)*. South-Western Thomson Learning.
- Dinardo, J., Johnston, J., 1996. *Econometric Methods*. McGraw Hill.
- Draper, D., 1995. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society Series B*, 57, 45 - 97, 1995.

- Draper, D., 1997. *On the Relationship Between Model Uncertainty and Inferential/Predictive Uncertainty*. Tech. Rept., University of Bath Statistics Group.
- Enders, W., 1995. *Applied Econometric Time-series*. J. Wiley and Sons.
- Gelman, A., Carlin, J., Stern, H., Rubin, D., 1995. *Bayesian Data Analysis*. Chapman & Hall/CRC.
- Gosling, J., Joy, B., Steele, G., Bracha, G., 2000. *The Java Language Specification, (2nd ed.)*. Addison Wesley Longman.
- Grewal, M. S., Andrews, A. P., 2001. *Kalman Filtering: Theory and Practice Using MATLAB (2nd ed.)*. J. Wiley and Sons.
- Harvey, A., 1994. *Forecasting, Structural Time-series Models and the Kalman Filter*. Cambridge University Press.
- Lee, E., 2002. *A Decision Process for Introducing New High-Technology Products*. Ph.D. Dissertation, Stanford University.
- Lilien, G., Kotler P., Moorthy K., 1992. *Marketing Models*. Prentice-Hall.
- Makridakis, S., Wheelwright, S., Hyndman, R., 1997. *Forecasting: Methods and Applications (3rd ed.)*. J. Wiley and Sons.
- Mentzer, J., Bienstock, C., 1998. *Sales Forecasting Management: Understanding the Techniques, Systems, and Management of the Sales Forecasting Process*. Sage Publications.
- Pole, A., West, M., 1989. Reference analysis of the DLM. *J. Time-series Analysis*, 10.
- Pole, A., West, M., Harrison, J. , 1994. *Applied Bayesian Forecasting and Time-series Analysis*. Chapman & Hall/CRC.
- Wald, A., 1947. *Sequential Analysis*. J. Wiley and Sons.
- West, M. , 1992. Approximating posterior distributions with mixtures. *J. Roy. Statist. Soc. (Ser. B)*, 55.
- West, M., Harrison, J., 1997. *Bayesian Forecasting and Dynamic Models (2nd ed.)*. Springer-Verlag.
- Wolfram, S., 1996. *The Mathematica Book (3rd ed.)*. Cambridge University Press.

About the Authors

Phillip M. Yelland is a senior staff engineer at Sun Microsystems Laboratories. He has worked in a wide variety of settings, ranging from formal methods research to software development management, both during his time at Sun and previously. Throughout, his activities have reflected a preoccupation with the application of theoretical results to the practical conduct of business, and his research currently centers on the use of statistical techniques for operations management. He has an M.A. and Ph.D. in Computer Science from the University of Cambridge in England, and an M.B.A. from the University of California at Berkeley.

Eunice Lee is a program manager within the World-Wide Operations organization at Sun Microsystems in Newark, California. She has over 15 years of industry experience in various roles at Boeing and Sun Microsystems; she has worked as a market planner, software engineer, systems analysts and consultant in Decision Analysis (DA). At Sun, she co-pioneered a new decision process, named KAPTUR, for forecasting product volumes of transitioning products using the techniques in DA. She holds a Ph.D. in Management Science and Engineering from Stanford University, an M.B.A from Claremont College, and a B.S. in Computer Science from UCI.