A Bayesian EWMA for Mean and Variance

by

Spencer Graves, Søren Bisgaard, and Murat Kulachi

Affiliations

Spencer Graves
Productive Systems Engineering
751 Emerson Ct.
San José, CA 95126
sgraves@prodsyse.com

Søren Bisgaard
Søren Bisgaard and Associates
PO Box 101
2315 B Forest Dr.
Annapolis, MD 21401
s.bisgaard@monitoring-intl.com

Murat Kulahci
Arizona State University
Dept. of Industrial Engineering
P. O. Box 875906
Tempe, AZ 85287
Murat.Kulahci@asu.edu

A Bayesian EWMA for Mean and Variance

ABSTRACT

We obtain simultaneous exponentially weighted moving averages (EWMAs) for mean and variance as the Bayesian prior for the next noisy observation of a normal random walk, where the information (reciprocal variance) of migration and noise have a common factor whose reciprocal follows a gamma-beta random walk. This procedure has applications in finance as an alternative to the popular ARCH / GARCH models, in addition to physical processes such as wear out and automotive misfire monitoring. The procedure is derived by applying Bayesian sequential updating (1. Observation, 2. Transition).

KEY WORDS: Exponentially weighted moving average; Bayesian Sequential Updating; On-board diagnostics (OBDs); autoregressive conditional heteroscedasticity (ARCH); misfire detection; wear out; fast initial response (FIR).

1. INTRODUCTION

This article describes a method for monitoring a process with changing mean and variance. An example appears in Figure 1. Somewhat similar images appear in financial and economics data and in data from many physical processes. This example presents the angular acceleration after 250 firing events in cylinder 4 in a V-8 engine. The plant (i.e., system monitored) was decelerating during the first half of the period shown here. The latter half shows the effect of requested acceleration combined with misfires: We see increases in both the variability from one observation to the next and in the apparent background rate of drift.

(Figure 1 about here)

Data on stock prices sometimes exhibit behavior similar to Figure 1: Surprising news about a company bursts upon investors, leading to a period of increased volatility in the price per share. Similar behavior is exhibited by many physical systems, including manufacturing processes. A piece of equipment ("plant") performs in a stable manner until a component crosses a deterioration threshold that degrades the consistency of performance of the plant, leading to an increase in both the short- and long-term variability. Figure 1 presents the angular acceleration accompanying 250 firing events for cylinder 4 measured at the front of a crankshaft of a V-8 engine. For almost the first half of the period portrayed here, the plant appears to be operating in a relatively stable mode with modest variability. Suddenly, we see a substantial jump followed by a period of elevated instability. During the initial, stable period, the vehicle is decelerating. Roughly 20% of the way through Figure 1, the engine begins misfiring occasionally on all cylinders due to inadequate spark. However, the misfires are not apparent in the figure until the throttle is opened roughly half way through Figure 1, whereupon the gap between complete and incomplete combustion generates the instabilities we see.

In this article, we will model these observations with Bayesian exponentially weighted moving averages (EWMAs) for mean and variance, discussed in Section 2. These are simple and elegant tools suitable for many applications where better modeling of variability may help improve understanding of the behavior of the plant. For misfire detection, this model might provide a bridge towards more sophisticated models that could aid in fault isolation [which cylinder(s) are misfiring] as well as mere detection of a misfire problem. The results of applying these tools to the data of Figure 1 appear in Figure 2. The EWMA for mean appears as the bold line in Figure 2.1, substantially smoothing the data. Two sets of dashed lines about this mean give 99.7% confidence bounds for the *mean* and for *the next observation*; these are Student's *t* confidence bounds with degrees of freedom reflecting the equivalent number of observations incorporated in the relevant inverse chi-square distribution of the EWMA for variance.

(Figure 2 about here)

Similarly, the dark solid line in Figure 2.2 presents the square root of the EWMA for variance of the predictive distribution. It starts large because the initial uncertainty about the mean at first use implies substantial uncertainty about the first observation. Information quickly accumulates about the location of the mean, which then allows information to accumulate about the variability as well. A pair of dashed lines close to the solid line give 99.7% confidence bounds for the uncertainty in the estimated predictive standard deviation, using as before the relevant inverse chi-square distribution of the EWMA for variance. A third dashed line gives a Student's *t* 99.7% upper bound for the absolute value of the prediction error.

We develop tools for this application using the two-step Bayesian sequential updating process (1. Observation and Bayes' theorem; 2. Transition and prior for the next observation) previously used to develop a Bayes-adjusted Cusum to detect an abrupt jump from a simple "good" to a simple "bad" hypothesis (Graves, Bisgaard and Kulahci 2002a), a Bayesian EWMA for mean only (Graves, Bisgaard and Kulahci 2002b) and more general Kalman filter monitors to isolate as well as detect problems (Graves et al. 2001). The previous Bayesian

EWMA (Graves, Bisgaard and Kulahci 2002b) assumes that the observation error variance is estimated from studies of gage repeatability and reproducibility, while the migration variance is estimated from reliability data. In contrast, we here assume that these variances are unknown and may change over time, although their ratio is assumed known. If the rate of change in the variance is zero, it provides a natural Bayesian foundation for estimating the system variance assuming a known ratio between the migration and observation variances.

This applies more general theory developed by West, Harrison, and Pole (West and Harrison 1999 and Pole, West and Harrison 1994) with two primary differences: First, they do not identify their evolving precision with an EWMA for variance; because of the widespread understanding of variances, standard deviations and EWMAs, we think this is a useful innovation. Second, they discuss the theory in terms of Student's *t* distributions, largely glossing over the normal-gamma origins of their Student's *t* distributions. We find this mode of reasoning difficult to follow, because Student's *t* distributions do not have the obvious convolution and Bayesian conjugate properties of normal distributions.

Therefore, Bayesian sequential updating is applied here to normal-gamma distributions in a cycle outlined briefly in Figure 3 and discussed in more detail in Section 2. Student's *t* confidence intervals for the means are then developed in Section 3 by integrating out a common gamma precision factor. A few sample computations are presented in Section 4, and concluding remarks appear in Section 5.

(Figure 3 about here)

The computations described below may appear more complicated than they really are. Applications requiring only smoothed estimates of mean and variability may need only EWMAs of mean and of squared prediction error (as a proportion of the relative predictive variance). If no fast initial response capability is needed, weights on the last observation and statistical control limits can be constructed with effort comparable to current EWMA procedures. The sample computations in Section 4 include several steps that merely involve renaming certain quantities. For example, the posterior mean at each point in time is numerically identical to the prior mean of the process at the arrival of the next observation and to the forecast mean for that observation, although confidence intervals associated with these three concepts are different. Maintaining a notational distinction helps us understand and discuss the different confidence intervals; it also facilitates applications to more general, e.g., Kalman filtering, situations where posterior, prior, and forecast means may be different.

Part of the complexity is due to the multitude of distributions involved: At each point in time, in addition to the basic migration and observation models, we have a predictive distribution and both prior and posterior distributions. To manage this complexity, we adopt a double subscript notation, e.g., $x_{t|t-1}$ for a prior and $x_{t|t}$ for a posterior mean for an unknown state x_t. [Harvey (1989) used a similar notational convention but not as extensively as we do here.]

2. BAYESIAN UPDATING WITH A DRIFTING MEAN AND VARIANCE

In this section, we develop the theory for a Bayesian exponentially weighted moving average (EWMA) for both mean and variance. The basic formulae are outlined in Table 1. This provides a standard EWMA for both mean and variance with the weight on the last observation changing with time, converging to an asymptote as information accumulates on the state of the process. As with the Bayesian EWMA for mean only, discussed by Graves, Bisgaard and Kulahci (2002b), this provides a very sensible fast initial response (FIR), adjusting the weight on the last observation to balance the relative information content of prior and observation. The theory follows naturally from the two-step Bayesian sequential updating procedure outlined in Figure 3.

(Table 1 about here)

Step 1. Observation and Bayes' Theorem. As each new observation arrives, the first step is to combine the information it contains about the condition of the plant (i.e., the process monitored) with the prior from previous experience. Then Step 2 will modify the resulting posterior to account for a transition in the plant anticipated before the next observation.

Step1.0. Model Assumptions. Specifically, at first use, the condition of the plant ($x_1 \mid f_1$) is assumed to follow $N(x_{1|0},\ s^2_{1|0}/f_1)$, where f_1 is assumed to follow a gamma distribution, $\Gamma(n_{1|0}/2,\ d_{1|0}/2)$. For future reference, we recall that $f(x_1, f_1) = f(x_1|f_1)f(f_1)$, where

$$f(x_1|f_1) \propto f_1^{1/2} \exp\left\{-\frac{f_1}{2}\left(\frac{x_1 - x_{1|0}}{s_{1|0}}\right)^2\right\}, \tag{1}$$

and

$$f(f_1) \propto f_1^{(n_{1|0}-2)/2} \exp\{-f_1 d_{1|0}/2\}, \tag{2}$$

so

$$f(x_1, f_1) \propto f_1^{(n_{1|0}-1)/2} \exp\left\{-\frac{f_1}{2}\left[\left(\frac{x_1 - x_{1|0}}{s_{1|0}}\right)^2 + d_{1|0}\right]\right\}. \tag{3}$$

(We often omit constants required to make probability densities integrate to one when they are not needed to understand what we are doing and may obscure our message.)

At each point in time, we observe

$$y_t = x_t + v_t, \text{ where } v_t \sim N(0,\ s_v^2/f_t),$$

so (4)

$$f\left(y_t \middle| x_t, f_t\right) \propto f_t^{1/2} \exp\left\{-\frac{f_t}{2}\left(\frac{y_t - x_t}{s_v}\right)^2\right\}.$$

Just before this observation, the prior for (x_t, f_t | D_{t-1}) is {$N(x_{t|t-1},\ s_{t|t-1}^2/f_t)$, $\Gamma(n_{t|t-1}/2,\ d_{t|t-1}/2)$}, where D_{t-1} = { y_{t-1}, y_{t-2},..., y_1, $x_{1|0}$, $s_{1|0}^2$, $n_{1|0}$, $d_{1|0}$}, the history available at time $t-1$:

$$f\left(x_t, f_t \middle| D_{t-1}\right) = f\left(x_t \middle| f_t, D_{t-1}\right) f\left(f_t \middle| D_{t-1}\right),$$

where

$$f\left(x_t \middle| f_t, D_{t-1}\right) \propto f_t^{1/2} \exp\left\{-\frac{f_t}{2}\left(\frac{x_t - x_{t|t-1}}{s_{t|t-1}}\right)^2\right\}, \tag{5}$$

and

$$f\left(f_t \middle| D_{t-1}\right) \propto f_t^{\left(n_{t|t-1}-2\right)/2} \exp\left\{-f_t d_{t|t-1}/2\right\}, \tag{6}$$

so

$$f\left(x_t, f_t \middle| D_{t-1}\right) \propto f_t^{\left(n_{t|t-1}-1\right)/2} \exp\left\{-\frac{f_t}{2}\left[\left(\frac{x_t - x_{t|t-1}}{s_{t|t-1}}\right)^2 + d_{t|t-1}\right]\right\}. \tag{7}$$

When t = 1, (5) - (7) is provided by the initial prior (1) - (3); later, it is provided by the output of Step 2, transition, after the previous observation.

As outlined in Figure 3, when a new observation arrives, this prior is converted to a posterior (Step 1), and the posterior is then modified to model a transition, producing a prior for the next observation (Step 2). We shall see that the resulting posterior and the new prior are both also normal-gamma, which we write as {$N(x_{t|t},\ s_{t|t}^2/f_t)$, $\Gamma(n_{t|t}/2,\ d_{t|t}/2)$} and {$N(x_{t+1|t},\ s_{t+1|t}^2/f_{t+1})$, $\Gamma(n_{t+1|t}/2,\ d_{t+1|t}/2)$}, respectively.

In particular, the transition in location is modeled as

$$x_{t+1} = x_t + w_t, \quad w_t \sim N(0,\ s_w^2/f_t). \tag{8}$$

Meanwhile, the transition for f_t involves discounting $\Gamma(n_{t|t}/2,\ d_{t|t}/2)$ to $\Gamma(n_{t+1|t}/2,\ d_{t+1|t}/2)$ for f_{t+1}, with $n_{t+1|t} = d\ n_{t|t}$ and $d_{t+1|t} = d\ d_{t|t}$, for some d with $0 < d \le 1$.

It seems appropriate at this point to discuss terminology. A squared reciprocal scale parameter such as f_t is sometimes called a precision (e.g., DeGroot 1970, p. 167; Bernardo and Smith 2000, pp. 121, 139). Bernardo and Smith (2000) discuss normal-gamma distributions like those we consider here but apparently without assigning names to parameters such as $s_{1|0}^2$ and f_1. In this article, at each point in time, the prior, posterior, observation, migration, and predictive distributions will all be normal-gamma. For t fixed, the parameter f_t will be the same in all of these distributions, though this parameter may change over time. To remind ourselves of this repeated use of f_t, we shall call it a "common precision factor"; by extension, we shall call its reciprocal f_t^{-1} a "common variance factor".

With this, we see that $s_{1|0}^2$ is the variance of x_1 as a proportion of the common variance factor f_1^{-1}; we will therefore call $s_{1|0}^2$ the "relative variance" of x_1 for short. We shall describe $s_{1|0}^{-2}$ as the "information" regarding x_1 as a proportion of the common precision factor f_1, or the "relative information" for short. We call this "relative information" rather than "relative precision" to emphasize the fact that the observed information, being the negative second derivative of the log(density) with respect to x_1, adds with Bayes' theorem (Graves 2002). This additive property of the observed information seems to provide greater insight for potentially non-normal observations or with non-linear applications than the concept of a "precision", being a squared reciprocal scale factor.

We now describe the use of Bayes' theorem in this context. To simplify the presentation of details, we break this activity into two substeps, preparing and updating.

Step 1.1. Preparing. We further divide "preparing" into three substeps: (1.1a) Predictive Distribution, (1.1b) Posterior Information and Variance, and (1.1c) Kalman Gain, as we now explain.

Step 1.1a. Predictive Distribution. We want the marginal distribution $(y_t \mid f_t,\ D_{t-1})$. From (4), we see that y_t is the sum of two independent, normally distributed random variables, so $(y_t \mid f_t,\ D_{t-1})$ is also a normal distribution, with mean and variance being the sums of the means and variances of x_t and v_t:

$$(y_t \mid f_t, D_{t-1}) \sim N(f_t,\ s^2_{y|t-1}/f_t),$$

where

$$f_t = x_{t|t-1}, \tag{9}$$

and

$$s^2_{y|t-1} = s^2_{t|t-1} + s^2_v . \tag{10}$$

[Note that f_t^{-1} is a common factor of the variances of $(x_t \mid f_t, D_{t-1})$ and the observation y_t per (4); it is therefore also a factor of the predictive variance.]

Step 1.1b. Posterior Information and Variance. Fisher's efficient score [derivative of the log density, $l(...)$] of the posterior is the prior score plus the score from the data:

$$\frac{\partial l(x_t \mid D_t)}{\partial x_t} = \frac{\partial l(x_t \mid D_{t-1})}{\partial x_t} + \frac{\partial l(y_t \mid x_t)}{\partial x_t} = \left[-\frac{x_t - x_{t|t-1}}{s^2_{t|t-1}/f_t}\right] + \left[-\frac{x_t - y_t}{s^2_v/f_t}\right] \tag{11}$$

(Graves 2002). From this, we see that the posterior score is linear in x_t with f_t as a factor and therefore can be written $\left[-f_t\left(x_t - x_{t|t}\right)/s_{t|t}^2\right]$ for appropriately chosen $x_{t|t}$ and $s_{t|t}^2$. The integral of this gives us the logarithm of the posterior density as a parabola up to an additive constant. Since the support of x_t runs over the entire real line, this proves that the posterior is $N(x_{t|t},\ s_{t|t}^2/f_t)$.

To determine $s_{t|t}^2$, we take another derivative of (11). This gives us "Bayes' Rule of Information" (Graves 2002) that the posterior (observed) information is the sum of the information from the prior and the data. Apart from the common precision factor f_t, this gives us the following:

$$s_{t|t}^{-2} = s_{t|t-1}^{-2} + s_v^{-2}. \tag{12}$$

Step 1.1c. Kalman Gain. We now let $x_t = 0$ in (11) and solve for $x_{t|t}$ as follows:

$$x_{t|t} = s_{t|t}^2\left\{s_{t|t-1}^{-2}x_{t|t-1} + s_v^{-2}y_t\right\}, \tag{13}$$

The weight on the last observation y_t in (13) is called the Kalman gain and will be denoted as follows:

$$K_t = s_{t|t}^2 s_v^{-2}. \tag{14}$$

To use K_t in (13), we first use (12) to obtain

$$s_{t|t-1}^{-2} = s_{t|t}^{-2} - s_v^{-2}.$$

We substitute this with (14) into (13) to get

$$\begin{aligned} x_{t|t} &= s_{t|t}^2\left\{\left(s_{t|t}^{-2} - s_v^{-2}\right)x_{t|t-1} + s_v^{-2}y_t\right\} \\ &= x_{t|t-1} + K_t\left(y_t - x_{t|t-1}\right). \end{aligned} \tag{15}$$

For plants with constant observation and transition variances, all the computations of substep 1.1 can be done offline except for the mean of the predictive distribution, f_t. With or without those offline computations, if these preparations are done prior to the arrival of the latest observation, y_t, it can shorten slightly the time required to update our knowledge of the state of the plant.

Step 1.2. Updating. In "updating", we (*a*) compute the prediction error and use that to compute (*b*) the posterior mean and (*c*) the posterior distribution of the common precision factor.

Step 1.2a. Prediction Error. When the observation y_t arrives, we compute the prediction error as,

$$e_t = y_t - f_t. \tag{16}$$

Step 1.2b. Posterior Mean. With the prediction error in hand, we multiply it by the Kalman gain and add the product to the prior mean to obtain the posterior mean per (15), as

$$x_{t|t} = x_{t|t-1} + K_t e_t . \tag{17}$$

Step 1.2c. Posterior Precision. We now combine the predictive distribution (9) - (10) with the prior for the common precision factor (6) as

$$f\left(y_t, f_t \middle| D_{t-1}\right) = f\left(y_t \middle| f_t, D_{t-1}\right) f\left(f_t \middle| D_{t-1}\right) \propto f_t^{\left(n_{t|t-1}-1\right)/2} \exp\left\{-\frac{f_t}{2}\left[\left(e_t/s_{y|t-1}\right)^2 + d_{t|t-1}\right]\right\}$$

$$\propto f_t^{\left(n_{t|t}-2\right)/2} \exp\left\{-f_t d_{t|t}/2\right\},$$

where

$$n_{t|t} = n_{t|t-1} + 1$$

and

$$d_{t|t} = \left(e_t/s_{y|t-1}\right)^2 + d_{t|t-1}. \tag{18}$$

But since $f(f_t \mid D_t) = f(f_t \mid y_t, D_{t-1}) = f(y_t, f_t \mid D_{t-1}) / f(y_t, \mid D_{t-1}) \propto f(y_t, f_t \mid D_{t-1})$, we see that $(f_t \mid D_t) \sim \Gamma(n_{t|t}/2, d_{t|t}/2)$.

This completes Step 1, observation, in Bayesian sequential updating as outlined in Figure 3. Next, we model the transition that we assume will occur before the next observation, per Step 2.

Step 2. Transition and Prior for the Next Observation. Given $(x_t \mid f_t, D_t)$ and $(f_t \mid D_t)$ from Step 1, we consider the transition for the common precision factor f_t to f_{t+1} and for the location x_t to x_{t+1}.

Step 2.1. Prior Precision. Following Pole, West and Harrison (1994), West and Harrison (1999), and Shephard (1994), we model a potential change in precision by discounting the chi-square / gamma degrees of freedom and scale factor by a constant d, with $0 < d \leq 1$, so

$$(f_{t+1} \mid D_t) \sim \Gamma(n_{t+1|t} / 2, d_{t+1|t} / 2),$$

where

$$n_{t+1|t} = d\, n_{t|t},$$

and

$$d_{t+1|t} = d\, d_{t|t}. \tag{19}$$

Pole, West and Harrison (1994, p. 61) describe this step by saying that, "no formal model is specified for scale evolution, the scale prior being directly defined as a discounted version of the previous posterior." West and Harrison (1999, p. 361) report that this particular variance discounting can be justified by assuming that

$$f_{t+1} = g_t f_t / d\ , \tag{20}$$

where $(g_t|D_t) \sim \text{Beta}[d\, n_{t|t}/2, (1 - d)n_{t|t}/2]$; for this distribution, $E(g_t|D_t) = d$, so $E(f_{t+1}|D_t) = f_t$.

It may help to understand d to note that [using (18) and (19)]

$$n_{t|t} \to 1/(1-d) \text{ as } t \to \infty$$

(West and Harrison, p. 362). However, since we are usually more interested in the future than the past, we substitute this into (19) to get the following:

$$n_{t+1|t} \to d/(1-d) \text{ as } t \to \infty \tag{21}$$

Thus, selecting d is equivalent to specifying the degrees of freedom in the steady-state chi-square distribution for the common precision factor f_t.

Expression (20) can be modified in a variety of ways to model, e.g., the increase in volatility of stock prices accompanying an increase in trading volume (e.g, Lamoureux and Lastrapes 1990) or the effect in Figure 1 of a change in throttle angle. However, with or without (20) and possible refinements, we must still ignore the difference between f_t and f_{t+1} in modeling the transition from x_t to x_{t+1}, which is part of Step 2.2.

Expressions (18) - (19) are equivalent to an EWMA for the common variance factor f_{t+1}^{-1}, which we define as $t_{t+1|t}^2 = d_{t+1|t}/n_{t+1|t}$. By (19) and (18), this is

$$t_{t+1|t}^2 = \frac{d\, d_{t|t}}{d\, n_{t|t}} = t_{t|t}^2 = \frac{d_{t|t-1} + \left(e_t/s_{y|t-1}\right)^2}{n_{t|t-1}+1} = \frac{n_{t|t-1} t_{t|t-1}^2 + \left(e_t/s_{y|t-1}\right)^2}{n_{t|t-1}+1},$$

so

$$t_{t+1|t}^2 = (1-l_t) t_{t|t-1}^2 + l_t \left(e_t/s_{y|t-1}\right)^2,$$

where (22)

$$l_t = 1/(n_{t|t-1} + 1) = 1/n_{t|t}.$$

We will use this to evaluate variability in Section 3 not conditioned on the unknown common precision factor f_t.

Step 2.2. Prior Mean and Variance. Given the posterior ($x_t \mid f_t, D_t$) from Step 1, with the transition (8), we get ($x_{t+1} \mid f_t, D_t$) ~ $N(x_{t+1|t},\ s^2_{t+1|t} / f_t)$, where

$$x_{t+1|t} = x_{t|t}, \tag{23}$$

and

$$s^2_{t+1|t} = s^2_{t|t} + s^2_w. \tag{24}$$

When we return to Step 1 for the next observation, we replace f_t with f_{t+1}. If d = 1, the common precision factor is assumed to be constant, so $f_{t+1} = f_t$, and this step is obvious. If d < 1, it is not clear (at least to the present authors) that our model assumptions necessarily imply that ($x_{t+1} \mid f_{t+1}, D_t$) ~ $N(x_{t+1|t},\ s^2_{t+1|t} / f_{t+1})$. However, even if it is not strictly true, it seems to be a reasonable approximation for many situations (at least with d close to 1), as witnessed by its use in multivariate state space applications discussed, e.g., by Pole, West and Harrison (1994) and West and Harrison (1999).

We now substitute (24) into (12) and the result into (14) to obtain the following recursion for K_t:

$$K_t^{-1} = \left\{ \left(r^2 + K_{t-1} \right)^{-1} + 1 \right\}, \tag{25}$$

where $r^2 = s^2_w / s^2_v$. [Graves, Bisgaard and Kulahci (2002b, sec. 5) discussed the behavior of K_t over time assuming $f_t = 1$. However, since f_t cancels, their analysis of K_t applies here as well.]

This completes Step 2. The resulting prior output from one point in time {$N(x_{t+1|t}$, $s^2_{t+1|t} / f_{t+1}$), $\Gamma(n_{t+1|t}/2,\ d_{t+1|t}/2)$} becomes an input prior for Step 1.1, {$N(x_{t|t-1},\ s^2_{t|t-1} / f_t)$, $\Gamma(n_{t|t-1}/2,\ d_{t|t-1}/2)$}, at the next point in time. In this way, observations are processed sequentially as they arrive. If the model (1) - (8) is correct, then the prior $\{N(x_{t+1|t},\ s^2_{t+1|t}/f_t),$ $\Gamma(n_{t+1|t}/2,\ d_{t+1|t}/2)\}$ summarizes all the information in $D_t = \{y_t, y_{t-1}, ..., y_1, x_{1|0}, s_{1|0}\}$ about the state of the plant at time $t+1$ [ignoring the approximation involved in replacing f_t by f_{t+1} following (23) - (24)].

The most important expressions in this section are summarized in Figure 4. Unfortunately, these results cannot be used directly, because most are conditioned upon the unknown common precision factor f_t. We next integrate out f_t, obtaining Student's t marginals (Section 3). The results are applied to the data of Figure 1 in Section 4 before a summary discussion in Section 5.

(Figure 4 about here)

3. STUDENT'S *t* CONFIDENCE INTERVALS

The results of Section 2 cannot be used directly, because they involve the unknown common precision factor f_t. To apply the results, we must integrate out f_t, obtaining Student's t marginals from normal-gamma distributions discussed in Section 2.

The normal-gamma density as in (3) and (7) is as follows:

$$f(x,f) \propto f^{(n-1)/2} \exp\left\{-\frac{f}{2}\left[\left(\frac{x-m}{s}\right)^2 + d\right]\right\}.$$

We integrate out f to get the following (Bernardo and Smith 2000, sec. 6.2.4):

$$f(x) \propto \left[\left(\frac{x-m}{s}\right)^2 + d\right]^{-(n+1)/2} \propto \left[1+\frac{(x-m)^2}{ds^2}\right]^{-(n+1)/2} \propto \left[1+\frac{(x-m)^2}{ns^2}\right]^{-(n+1)/2},$$

where

$$s^2 = \left(s\ ^2 d/n\right) = s\ ^2 t\ ^2 \text{ and } t\ ^2 = d/n.$$

We summarize this by observing that x has a Student's t distribution with n degrees of freedom and with center and scale of m and s,

$$x \sim t(m, s^2; n).$$

Applying this to the prior (7) gives us

$$(x_t \mid D_{t-1}) \sim t(x_{t|t-1},\ s^2_{t|t-1};\ n_{t|t-1}),$$

with (26)

$$s_{t|t-1} = s_{\ t|t-1}\, t_{t|t-1},$$

and $t^{\ 2}_{t|t-1}$ is the common variance factor EWMA (22). This was used in Figure 2.1 to determine the inner set of dashed lines as a 99.7% confidence interval for $\left(x_t \mid D_{t-1}\right)$.

Similarly, for the predictive distribution not conditioned on the unknown precision $f_{\ t}$, we get the following from (9) - (10) and (6):

$$(y_t \mid D_{t-1}) \sim t\left(f_t, s^2_{y|t-1}; n_{t|t-1}\right),$$

with (27)

$$s_{y|t-1} = s_{\ y|t-1}\, t_{t|t-1}.$$

This was used to compute the outer dashed lines in Figures 2.1 and 2.2. The inner pair of dashed lines in Figure 2.2 utilize the (0.0015 and 0.9985) quantiles of the gamma distribution of (19) to place confidence limits on $s_{y|t-1}$.

We now discuss the computation of a few of the numbers plotted in Figure 2.

4. SAMPLE COMPUTATIONS

The sample computations described in this section are more complex than required for many applications. For example, the forecast for the next observation (f_t), the current prior mean $(x_{t|t-1})$, and the previous posterior mean $(x_{t-1|t-1})$ are conceptually three different quantities that are numerically equal in the present context. The distinction is visible in the different confidence intervals associated with the three concepts. In Figure 2, dashed lines represent confidence intervals based on the forecast and the prior. The confidence intervals associated with the posterior are slightly narrower than the confidence intervals for the prior and are not shown because we are more concerned with the future than the past (at least for Figure 2). This careful distinction in notation has helped us understand the EWMA and generalizations to, for example, Kalman filtering, where f_t, $x_{t|t-1}$, and $x_{t-1|t-1}$ may all be distinct. These distinctions are maintained in this section, though they were suppressed in Table 1.

In the present model, $\mathrm{var}(y_t \mid x_t, f_t) = s_v^2/f_t$ per (4) and $\mathrm{var}(x_t \mid x_{t-1}, f_t) = s_w^2/f_t$ per (8) [ignoring the change from f_{t-1} to f_t discussed following (24)]. Since there are no other constraints on f_t, we can without loss of generality set $s_v^2 = 1$. With this choice, f_t becomes the observation precision, so $t_{t|t-1}^2$ estimates the observation variance, and s_w^2 = $\mathrm{var}(x_t \mid f_t)/\mathrm{var}(y_t \mid f_t)$ = r^2 = the migration variance as a proportion of the observation variance.

With this choice, Table 2 begins by recording that the relative observation variance s_v^2 and information s_v^{-2} are both 1. Similarly, Table 2 reports that the relative migration variance

s_w^2 is assumed to be 0.01, while the variance discount factor d = 0.98. As discussed by Graves, Bisgaard and Kulahci (2002b), the migration variance $s_w^2 = r^2$ is related to reliability and is equivalent to specifying the degree of smoothing. Meanwhile, per (21), d = 0.98 corresponds to an asymptotic degrees of freedom of $n_{\infty+1|\infty}$ = 49, roughly one fifth of the observations in Figures 1 and 2.

(Table 2 about here)

The first row in the body of the table gives the first three observations y_t plotted in Figure 1. The initial prior for the mean per (1) is specified in terms of the mean $x_{1|0}$ and standard deviation $s_{1|0}$. A rough estimate obtained simply by looking at Figure 1 is $x_{1|0}$ = 0 and $s_{1|0}$ = 25; this latter number means that $s_{1|0}^2$ = 625 and $s_{1|0}^{-2}$ = 0.0016. Better estimates could be obtained if the application justified the extra expense. This follows because monitors are only designed where previous experience suggests the possibility of problems (Graves, Bisgaard and Kulahci 2002b). This experience provides an objective external reference population, which could be accessed to provide better estimates of parameters such as $x_{1|0}$ and $s_{1|0}$.

For later observations, $x_{t|t-1}$, $s_{t|t-1}^2$, and $s_{t|t-1}^{-2}$ are taken from Step 2.2 at the bottom of Table 2. We carry both relative variance and information parameters in Table 2 because a sum of independent random variables requires addition of variances, while "Bayes' Rule of Information" (Graves 2002) tells us to add the (observed) information.

Similarly, the initial value of the EWMA for variance $t_{1|0}^{2}$ was chosen as 3^2 just by eying Figure 1. It was assigned 1 degree of freedom ($n_{1|0}$) to indicate that we are assuming this method of estimation is roughly equivalent to one single good number. As for $x_{1|0}$ and $s_{1|0}$, better estimates for $t_{1|0}^{2}$ and $n_{1|0}$ could be obtained if the application justified the effort.

Thus, $t_{1|0}^{2} = 9$. For $t > 1$, we get $t_{t|t-1}^{2}$ and $n_{t|t-1}$, from Step 2.1 at the bottom of Table 2. Next, the sample standard deviation for x_t, $s_{t|t-1}$, is computed per (26) as the square root of the product of the EWMA for variance and the relative prior variance per (26), producing $s_{1|0}$ = 75. We compute a confidence interval about $x_{t|t-1}$ using a Student's t distribution with $n_{t|t-1}$ degrees of freedom with scale factor $s_{t|t-1}$. At $t = 1$, we have 1 degree of freedom, which produces 212.2 as the 0.9985 quantile of the relevant Student's t distribution. This times $s_{1|0}$ = 75 is 15,915; we add and subtract this from $x_{1|0} = 0$ to get the corresponding confidence limits in Table 2. This produces the inner pair of dashed lines in Figure 2.1.

With the prior specified for each step, we now proceed as outlined in Figure 4 to write down the predictive distribution, Step 1.1a. The predictive mean f_t is copied from the prior mean $x_{t|t-1}$, and the relative predictive variance $s_{y|t-1}^{2}$ is the sum of the relative prior and observation variance parameters, which produces 626 for the first observation. The square root of the product of the EWMA for variance $t_{t|t-1}^{2}$ and the relative predictive variance $s_{y|t-1}^{2}$ gives us the predictive sample standard deviation $s_{y|t-1}$. For $t = 1$, this is 75.06, slightly larger than $s_{1|0}$; after $t = 1$, the predictive sample standard deviation $s_{y|t-1}$ is noticeably larger than $s_{t|t-}$

$_1$, because the prior quickly becomes more informative than a single observation; for $t = 1$, the opposite is true. The predictive sample standard deviation is the solid line in Figure 2.2.

To get a 99.7% tolerance interval for the new observation and for the absolute prediction error, we repeat the same logic as for the confidence interval for the prior mean. This gives us the outer set of dashed lines in Figures 2.1 and 2.2. A confidence interval for the predictive sample standard deviation is obtained by referring it to a chi-square distribution with $n_{t|t-1}$ degrees of freedom. This produces the inner pair of dashed lines in Figure 2.2.

We now proceed to Step 1.1b, computing the relative posterior information as the sum of the relative information from prior and observation. The relative posterior variance is the reciprocal of the corresponding relative information. Next, in Step 1.1c we compute the weight on the last observation, the Kalman gain, which per (14) is the relative posterior variance times the relative information from the observation. For the first observation, this is 0.998, reflecting the fact that the first observation is substantially more informative than our barely informative prior. If the prior had been completely non-informative, the relative prior information would have been exactly 0, in which case the Kalman gain would have been exactly 1. For $t = 2$, the Kalman gain is 0.502. Even though the posterior from $t = 1$ is slightly more informative than a single observation, the migration with variance parameter $s_w^2 = 0.01$ makes the prior for $t = 2$ slightly less informative than a single observation. Similarly, the Kalman gain for the third observation is slightly greater than 1/3; with $r^2 = s_w^2 / s_v^2 = 0.01$, the Kalman gain continues down to an asymptote at 0.0951, which we get from letting $K_t = K_{t-1} = K_\infty$ in (25) (Graves, Bisgaard and Kulahci 2002b, sec. 5). This completes the preparations that could potentially be performed in real-time applications before the observation actually arrived.

With the new observation in hand, we first compute the prediction error e_t per (16), Step 1.2a. This is used to update the EWMAs for both mean and for variance. To prepare for updating the EWMA for variance, we square this and divide by its corresponding relative variance, obtaining 0.468 for $t = 1$.

We also include here the Student's t log(likelihood). This is not needed when computing only one EWMA in isolation. However, there are many uses for likelihood. For example, West and Harrison (1999, sec. 11.4.2) recommend the use of Bayes' factors for evaluating one model relative to another; their use of Bayes' factors is essentially equivalent to a traditional one-sided cumulative sum of log(likelihood ratio) or to the Bayes-adjusted Cusum of Graves, Bisgaard and Kulahci (2002a). For many applications, an appropriate likelihood rests on the marginal predictive distribution for the next observation, after integrating out the unknown common precision factor f_t.

Step 1.2b, (17), tells us to multiply the prediction error by the Kalman gain and add to the prior mean to get the posterior mean. Similarly, in Step 1.2c, we add 1 to the prior degrees of freedom to get the posterior degrees of freedom, which is 2 for $t = 1$. We next compute the weight on the last standardized squared prediction error as the reciprocal of the posterior degrees of freedom, per (22). This gives us 0.5 for the first observation, which is consistent with our assumption that the prior has the information content of one observation ($n_{1|0} = 1$). We use these numbers to complete the computation of the posterior EWMA for variance, obtaining $t^2_{t|t}$ = 4.734 for t = 1. This completes Step 1.

It remains to modify the posterior to account for anticipated migration prior to the next observation. Per (22), $t^2_{t+1|t} = t^2_{t|t}$. However, the degrees of freedom are discounted by the factor d. For t = 1 with d = 0.98 this discounts the posterior degrees of freedom from 2 to 1.96 for the prior at t = 2. Per (23), the future prior mean $x_{t|t-1}$ for t = 2 is equal to the present posterior mean $x_{t+1|t} = x_{t|t} = (-17.081)$ for t = 1. Similarly, the future relative prior variance is the relative posterior plus migration variances, giving us 1.008 for t = 1. We reciprocate this to get the relative prior information. For reference, we also compute the corresponding sample standard deviation as the square root of the product of the prior EWMA for variance and the relative prior information, which is $s_{t+1|t}$ = 2.185 for t = 1.

5. DISCUSSION

We believe that this discussion of a Bayesian EWMA for mean and variance illustrates the power of Bayesian sequential updating as a general principle for designing monitors; for other applications of this principle, see Graves, Bisgaard and Kulahci (2002a, b) and Graves et al. (2001). Other procedures for monitoring mean and variance have previously appeared in the literature, but without such obvious ties to a unifying principle for monitor design. For example, Gan (1995) compared four schemes proposed for simultaneous monitoring of center and variability. These included a Cusum and EWMAs of powers of observations and log(standard deviation). It would, of course, be interesting to extend Gan's study to include the scheme considered here. Beyond this, we suspect that Bayesian sequential updating considering various non-normal distributions might produce monitoring schemes reasonably well approximated by the alternatives Gan considered. Such research could help in two ways. First, it could help people design monitors based on data analysis suggesting alternative distributions for observations and transitions, in the spirit of Box (1980) and Chen and Box (1990). Second, it would help further the development of a general theory for monitor design. We shall leave this for future research.

We suspect that expression (20)

$$\log(f_{t+1}) = \log(f_t) + \log(g_t/d)$$

provides a fertile foundation for generalizations, modeling the impact on this common precision factor of exogenous variables, in the spirit of generalized autoregressive conditional heteroscedasticity (GARCH). With or without these generalizations, we feel that further research is needed to understand the extent of the approximation involved in the apparently ad hoc transition from $(x_{t+1} \mid f_t, D_t) \sim N(x_{t+1|t}, s^2_{t+1|t}/f_t)$ to $(x_{t+1} \mid f_{t+1}, D_t) \sim N(x_{t+1|t}, s^2_{t+1|t}/f_{t+1})$. This approximation has worked well for West and Harrison (1999) and Pole, West and Harrison (1994). However, it is not clear (at least to the present authors) how well this will work with rapid changes in the common precision factor f_t.

It is relatively straightforward to generalize this work to multivariate state spaces and observations with possibly non-normal observations and nonlinear observation and transition relationships, *provided* (*a*) normal distributions provide adequate approximations for both prior and posterior and (*b*) first order Taylor approximations can be used in the standard ways. In these situations, the common precision factors, f_t, will still be scalars (see, e.g., West and Harrison 1999, sec. 4.6), as will be the accompanying EWMA for variance. It may be possible to use a multivariate normal - Wishart distribution in a similar way (Bernardo and Smith 2000, sec. 3.2.5), but this could raise other questions of parsimonious modeling.

6. ACKNOWLEDGEMENTS

The work reported here was funded in part by the Low Emissions Research and Development Partnership of DaimlerChrysler, Ford and General Motors. The authors wish to thank John Van Gilder and Tom Ting of General Motors, Ken Marko and John James of Ford, and Hal Zatorski and Cuiping Wu of DaimlerChrysler for many stimulating discussions about this and related monitoring issues.

REFERENCES

Bernardo, J. M., and A. F. M. Smith (2000) *Bayesian Theory* (NY: Wiley).

Box, G. E. P. (1980) “Sampling and Bayes' Inference in Scientific Modelling and Robustness” (with discussion), *Journal of the Royal Statistical Society*, Series A, 143: 383-430.

Chen, G. G., and Box, G. E. P. (1990) “The Weighting Pattern of a Bayesian Robust Estimator”, pp. 3-21 in L. D. Kenneth adn J. L. Arthur (eds.) *Robust Regression: Analysis and Applications* (NY: Marcel Dekker)

DeGroot, M. H. (1970) *Optimal Statistical Decisions* (NY: McGraw-Hill).

Gan, F. F. (1995) "Joint Monitoring of Process Mean and Variance Using Exponentially Weighted Moving Average Control Charts", *Technometrics*, 37: 446-453.

Graves, S. B. (2002) "Bayes' Rule of Information" (unpublished technical report downloadable from www.prodsyse.com).

Graves, S. B., Bisgaard, S., and Kulahci, M. (2002a) "A Bayes-Adjusted Cusum" (unpublished technical report downloadable from www.prodsyse.com).

Graves, S. B., Bisgaard, S., and Kulahci, M. (2002b) "Designing Bayesian EWMA Monitors Using Gage R & R and Reliability Data" (unpublished technical report downloadable from www.prodsyse.com).

Graves, S. B., Bisgaard, S., Kulahci, M., Van Gilder, J., Ting, T., Marko, K., James, J., Zatorski, H., and Wu, C. (2001) "Foundations of Monitoring Dynamic Systems" (unpublished technical report downloadable from www.prodsyse.com).

Harvey, A. C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter* (NY: Cambridge University Press).

Lamoureux, C. G., and Lastrapes, W. D. (1990) "Heteroskedasticity in Stock Return Data: Volume versus GARCH Effects", *Journal of Finance*, 45, pp. 221-229.

Pole, A., West, M., and Harrison, H. (1994) *Applied Bayesian Forecasting and Time Series Analysis* (NY: Chapman & Hall)

Shephard, N. (1994) "State Space Alternatives to Integrated GARCH Processes", *Journal of Econometrics*, 60, 181-202.

West, M. and Harrison, P. J. (1999) *Bayesian Forecasting and Dynamic Models*, 2nd ed., corrected 2nd printing (NY: Springer).

Figure 1. A Time Series with Changing Mean and Variability

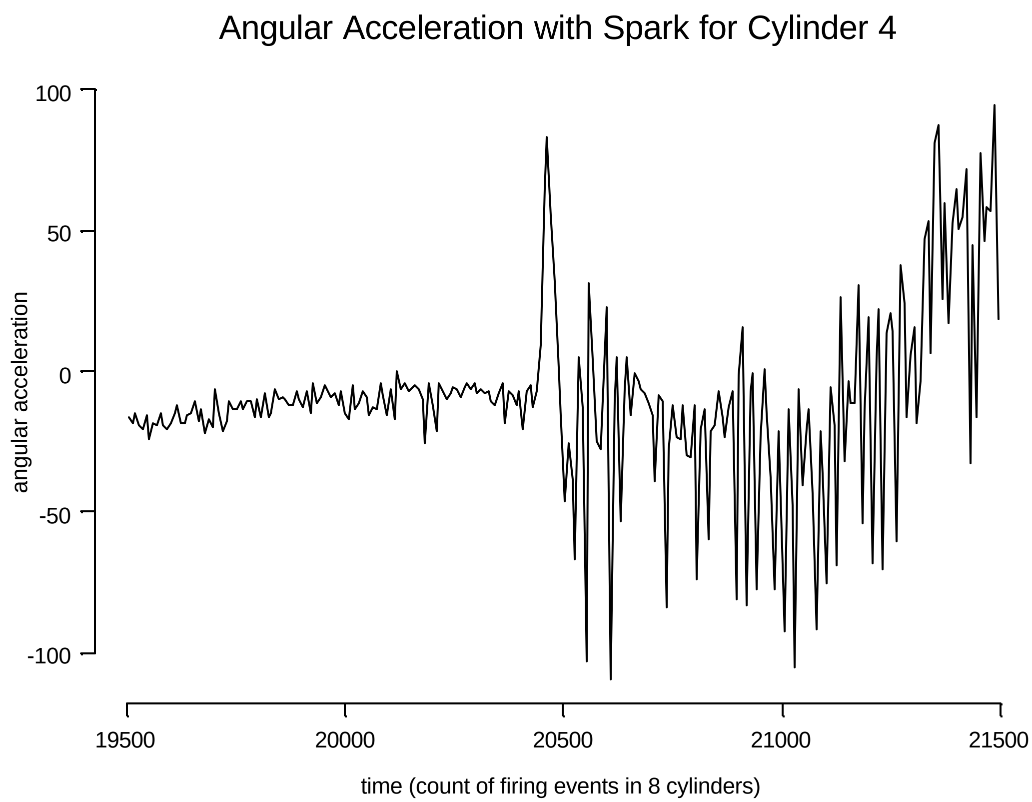

Figure 2. Example Smoothing of Mean and Standard Deviation

Figure 2.1. Data and Drifting Mean

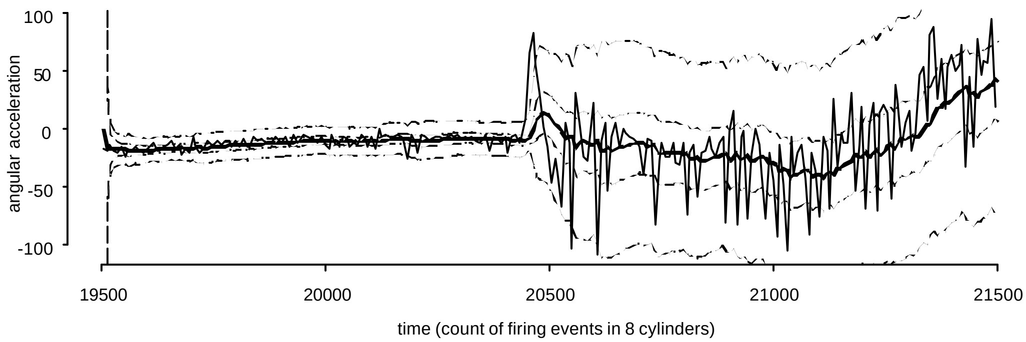

Figure 2.2. Absolute Prediction Error and Smoothed Standard Deviation of Prediction Error

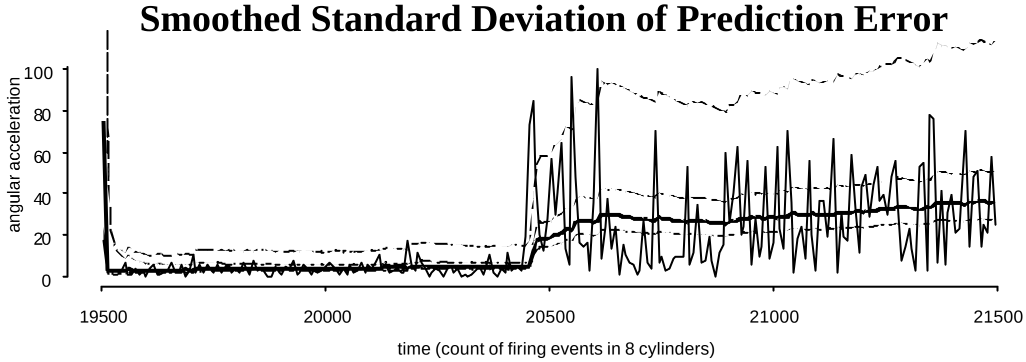

Figure 3. Bayesian Sequential Updating of Mean and Variance

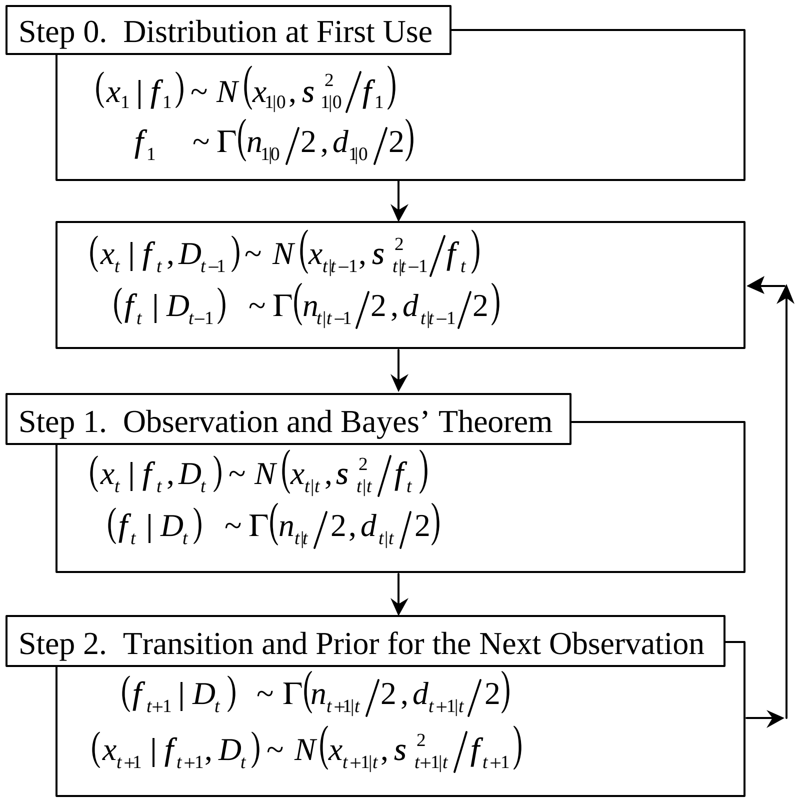

Figure 4. Bayesian EWMA Normal-Gamma Iteration

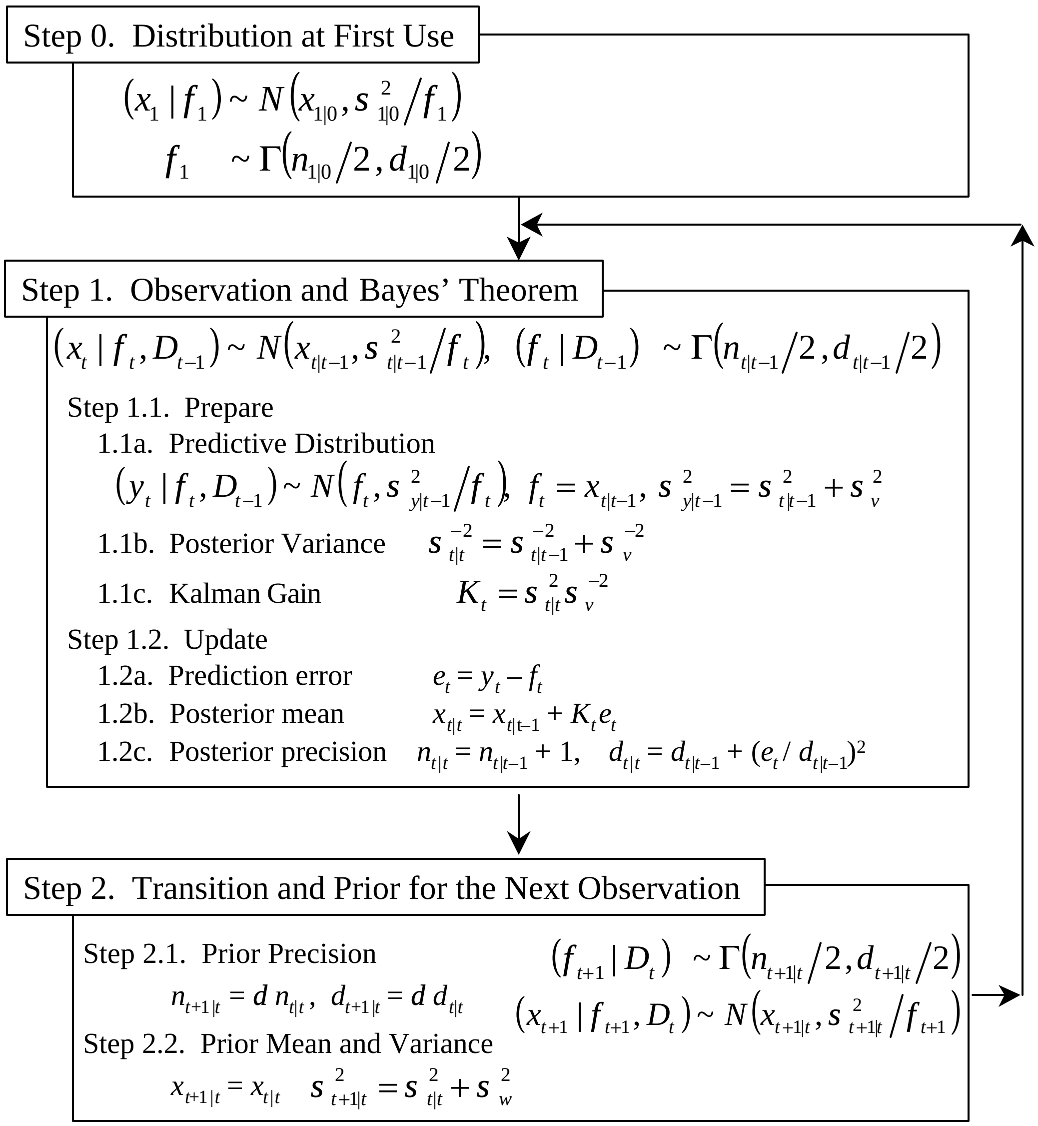

Table 1. Algorithm for Bayesian EWMA for Mean and Variance

Assume		
Observation	$y_t = x_t + v_t, \;\; v_t \sim N\left(0, s_v^2 / f_t\right)$	(4)
Migration	$x_{t+1} = x_t + w_t, \;\; w_t \sim N\left(0, s_w^2 / f_t\right)$	(8)
EWMA $x_{t\mid t-1}$ for mean x_t		
	$x_{t+1\mid t} = \left(1 - K_t\right) x_{t\mid t-1} + K_t y_t = x_{t\mid t-1} + K_t e_t$	(17) & (23)
Prediction error	$e_t = y_t - x_{t\mid t-1}$	(16)
Kalman gain	$K_t = 1 / \left\{1 + 1 / \left(r^2 + K_{t-1}\right)\right\}, \;\; r^2 = s_w^2 / s_v^2$	(25)
EWMA $t_{t\mid t-1}^2$ for the common variance factor f_t^{-1}		
	$t_{t+1\mid t}^2 = \left(1 - l_t\right) t_{t\mid t-1}^2 + l_t \left(e_t / s_{y\mid t-1}\right)^2$	(22)
Weight on the last prediction error	$l_t = 1 / \left(n_{t\mid t-1} + 1\right)$	
c^2 / Student's *t* degrees of freedom	$n_{t+1\mid t} = d\left(n_{t\mid t-1} + 1\right), \;\; 0 < d \leq 1$	(19) & (18)
Relative predictive error variance	$s_{y\mid t-1}^2 = s_{t\mid t-1}^2 + s_v^2$	(10)
Relative prior variance	$s_{t+1\mid t}^2 = \left(s_{t\mid t-1}^{-2} + s_v^{-2}\right)^{-1} + s_w^2$	(24) & (12)
Confidence interval for x_t via Student's *t*		
	$\left(x_t \mid D_{t-1}\right) \sim t\left(x_{t\mid t-1}, s_{t\mid t-1}^2; n_{t\mid t-1}\right)$	
Sample variance for the mean	$s_{t\mid t-1}^2 = s_{t\mid t-1}^2 t_{t\mid t-1}^2$	(26)
Prediction error via Student's *t*		
	$\left(e_t \mid D_{t-1}\right) \sim t\left(0, s_{y\mid t-1}^2; n_{t\mid t-1}\right)$	
Sample variance for prediction error	$s_{y\mid t-1}^2 = s_{y\mid t-1}^2 t_{t\mid t-1}^2$	(27)

Table 2. Illustrative Calculations for Bayesian Sequential Updating

Observation variability:

(4) relative variance $s_v^2 = 1$; relative information $s_v^{-2} = 1$

Transition:

(8) relative migration variance $s_w^2 = 0.01$; (19) variance discount factor $d = 0.98$

Step

		1	2	3
	time	1	2	3
1. Observation and Bayes' Theorem				
Observation: Measured angular acceleration	$y_t =$	-17.108	-19.095	-14.985
1.0. Prior (1) - (3), (5) - (7)				
mean	$x_{t\|t-1} =$	0.000	-17.081	-18.092
relative information	$s^{-2}_{t\|t-1} =$	0.0016	0.992	1.953
variance	$s^{2}_{t\|t-1} =$	625.000	1.008	0.512
EWMA for variance	$t^{2}_{t\|t-1} =$	9.000	4.734	3.817
degrees of freedom	$n_{t\|t-1} =$	1.000	1.960	2.901
sample standard deviation (26)	$s_{t\|t-1} =$	75.000	2.185	1.398
99.7% confidence interval				
Student's t (a = 0.0015)		212.205	19.080	9.316
for the mean	upper limit	15915.35	24.606	-5.067
for the mean	lower limit	-15915.35	-58.767	-31.117
1.1. Prepare				
1.1a. Predictive distribution				
mean (9)	$f_t =$	0.000	-17.081	-18.092
relative variance (10)	$s^{2}_{y\|t-1} =$	626.000	2.008	1.512
sample standard deviation (27)	$s_{y\|t-1} =$	75.060	3.083	2.402
99.7% confidence interval				
Student's t (a = 0.0015)		212.205	19.080	9.316
for the observation	upper limit	15928.10	41.750	4.290
for the observation	lower limit	-15928.10	-75.912	-40.474
for (absolute prediction error $\|e_t\|$) =		15928.10	58.831	22.382
for sample standard deviation $s_{y\|t-1}$				
	$s^2_{0.0015} = c^2\left(0.0015; n_{t\|t-1}\right)/n_{t\|t-1} =$	3.53×10^{-6}	1.33×10^{-3}	1.54×10^{-3}
	$s^2_{0.9985} = c^2\left(0.9985; n_{t\|t-1}\right)/n_{t\|t-1} =$	10.079	6.582	6.487
	upper limit $s_{y\|t-1}/s_{0.0015}$	39926.11	84.550	61.279
	lower limit $s_{y\|t-1}/s_{0.9985}$	23.643	1.202	0.943

1.1b. Posterior variability (12)				
relative information	$s_{t\|t}^{-2}$ =	1.0016	1.992	2.953
variance	$s_{t\|t}^{2}$ =	0.998	0.502	0.339
1.1c. Kalman gain (14)	K_t =	0.998	0.502	0.339
1.2. Update				
1.2a. Prediction error (16)	e_t =	-17.108	-2.014	3.107
standardize squared prediction error	$\left(e_t^2 / s_{y\|t-1}^{2}\right)$ =	0.468	2.020	6.384
log(likelihood)		-5.514	-2.460	-2.768
1.2b. Posterior mean (17)	$x_{t\|t}$ =	-17.081	-18.092	-17.040
1.2c. Posterior common precision factor				
degrees of freedom (18)	$n_{t\|t}$ =	2.000	2.960	3.901
weight on squared prediction error (22)	l_t =	0.500	0.338	0.256
deviation of standardized squared prediction error from prior relative variance	$\left[\left(e_t^2 / s_{y\|t-1}^{2}\right) - t_{t\|t-1}^{2}\right]$ =	-8.532	-2.713	2.567
EWMA for variance (22)	$t_{t\|t}^{2}$ =	4.734	3.817	4.475

2. Transition and prior for the next observation

2.1. Prior precision				
EWMA for variance (22)	$t_{t+1\|t}^{2}$ =	4.734	3.817	4.475
degrees of freedom (19)	$n_{t+1\|t}$ =	1.960	2.901	3.823
2.2. For process average				
mean (23)	$x_{t+1\|t}$ =	-17.081	-18.092	-17.040
relative variance (24)	$s_{t+1\|t}^{2}$ =	1.008	0.512	0.349
information	$s_{t+1\|t}^{-2}$ =	0.992	1.953	2.868
sample standard deviation	$s_{t+1\|t}$ =	2.185	1.398	1.249

Figure Captions

Figure 1. A Time Series with Changing Mean and Variability

Figure 2. Example Smoothing of Mean and Standard Deviation

Figure 3. Bayesian Sequential Updating of Mean and Variance

Figure 4. Bayesian EWMA Normal-Gamma Iteration