Foundations of Monitoring Dynamic Systems

Spencer Graves, Søren Bisgaard, Murat Kulachi, John Van Gilder,
Tom Ting, Ken Marko John James, Hal Zatorski, and Cuiping Wu

Bayesian Sequential Updating:

Step 1. Observation: Updating knowledge with Bayes' theorem

Step 2. Deterioration: Adjusting the posterior to produce the prior for the next observation

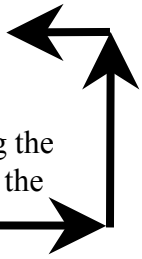
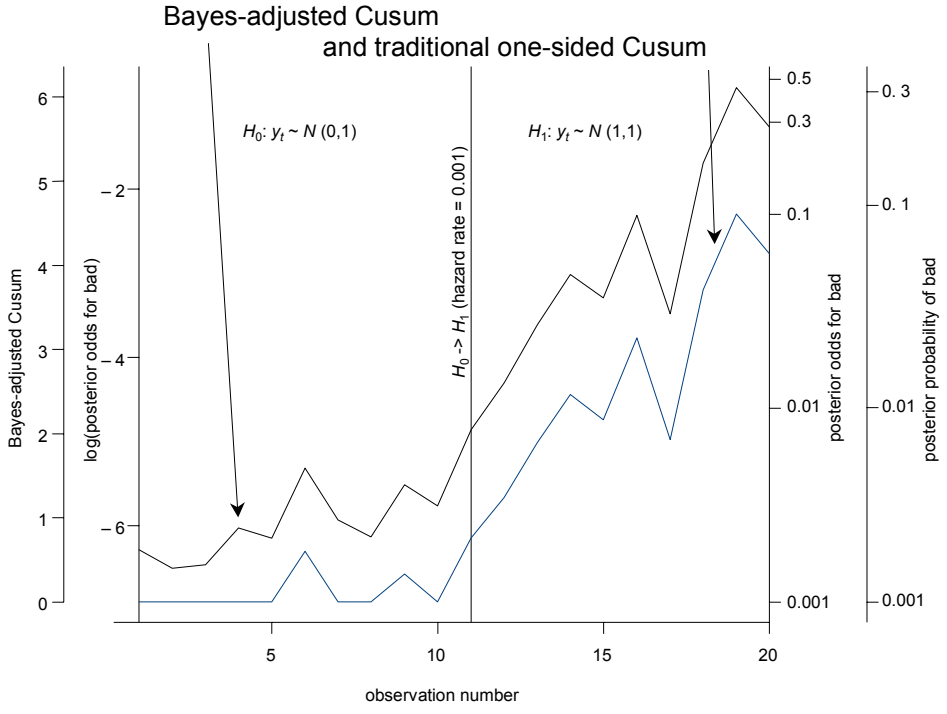


Figure 2.1. One-Sided and a Bayes-Adjusted Cusum Simulations



It is sometimes more important to isolate a fault than identify it.

Foundations of Monitoring Dynamic Systems

Spencer Graves, Søren Bisgaard, Murat Kulachi, John Van Gilder,
Tom Ting, Ken Marko John James, Hal Zatorski, and Cuiping Wu

Foundations of Monitoring Dynamic Systems

CONTENTS

PREFACE

Comments are invited on this prepublication edition of "Foundations of Monitoring Dynamic Systems". Your questions and suggestions may help improve future written and oral presentations of these ideas. Please email sgraves@prodsyse.com. Thanks.

ABSTRACT

The use of sophisticated computer controls of systems provides growing opportunities for designing additional systems for malfunction detection. Prime examples are provided by legally mandated On-Board Diagnostics to detect malfunctions in the emission controls on new automobiles sold today in the US, Canada and Europe, but the techniques can be profitably applied in many situations where data are collected over time. This article asserts that Bayesian sequential updating provides a general, unifying principle for understanding and designing monitors, i.e., systems that in real time can monitor, detect and isolate malfunctions. Bayesian sequential updating includes as special cases exponentially weighted moving averages (EWMAs) and Kalman filters more generally. In other applications, the Bayesian sequential monitor is accurately approximated by Page's one-sided Cusum of log(likelihood ratio).

KEY WORDS: Monitoring; Bayesian Sequential Updating; Cusum of log(likelihood ratio); Kalman Filtering; exponentially weighted moving average (EWMA); Bayes-adjusted Cusum; statistical process control (SPC).

1. INTRODUCTION

With the increasing computerization of products from simple to highly complex, the opportunities and demands for real-time diagnostic monitoring systems is growing rapidly. Prime examples are provided by On-board diagnostics (OBDs) to detect malfunctions in the emission controls required by law in new automobiles sold in the US, Canada and Europe. Other examples include modern heart pacemakers and implantable defibrillators that monitor both the patient and the device itself: They monitor the patient's condition and intervene only when necessary, and they sound an audible alarm if the electrical leads are corroded or the battery is low. Further examples help isolate opportunities for process improvement in complex manufacturing processes.

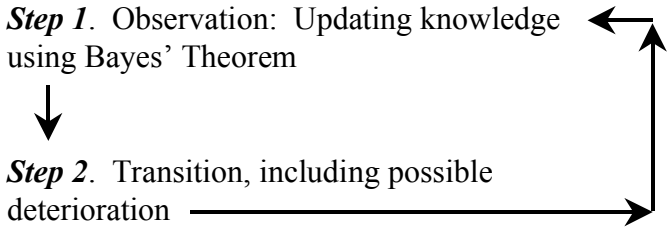
Monitors for complex systems (here called "plants" for consistency with the control theory literature) can be designed to isolate the specific component or subsystem subject to an actual or impending malfunction, making repairs easier, more certain, and less costly. In manufacturing, model-based monitors can be designed to compare the performance of different pieces of equipment ostensibly doing the same thing and alert appropriate personnel when a substantive difference is detected.

Monitoring is different from testing. The difference can be seen in clinical trials, which do both. The primary purpose of a clinical trial is to test a new therapy for safety and effectiveness; this is a property of the therapy that is not expected to change over time. Monitoring on the other hand is intended for systems that may start out good and then later change. In clinical trials, the condition of patients is monitored for indications that the treatment should be altered or discontinued. Statistical tests are evaluated on the basis of the probabilities of errors of Types I and II, false alarms and failures to detect.

Monitors, however, are more than just repetitive tests. Their design criteria should be specified more in terms of the average delay to detection or the probability of detection in a certain period of time, balanced by, for example, the probability of a false alarm in the design life of the "plant" being monitored; see Box et al. (2000) or Box et al. (2002).

In this report, we describe monitoring in terms of Bayesian sequential updating, which we define as a two-step iteration portrayed in Figure 1.1: (1) observation and (2) transition. If one is concerned with detecting single outlying observations, there is no transition step, and statistical control charts are appropriate. This is one extreme of the general rule that the nature of the transition step in large part drives the choice of a monitor.

Figure 1.1. Bayesian Sequential Updating



In particular, if the transition occurs as an abrupt jump from one state to another, Bayesian sequential updating produces a monitor that in most applications is quite similar to a cumulative sum (Cusum) of log(likelihood ratio), as we explain in section 2. In so

doing, we introduce a new recursion for the Girshick-Rubin (1952) Bayesian monitor that works with non-i.i.d. observations and non-constant hazard.

On the other hand, if the transition occurs as a gradual migration (normal random walk) in the mean of a univariate series of observations, Bayesian sequential updating gives us a new fast initial response (FIR) strategy for an exponentially weighted moving average (EWMA), as we explain in section 3. We also establish connections between (*a*) the initial distribution and some objective reference distribution (based, e.g., on experience with comparable manufacturing or comparable clinical trials) on the one hand, and (*b*) the migration rates and lifetime distributions on the other. To our knowledge, this is the first discussion in the literature of these objective foundations for the priors in Bayesian sequential updating. This permits appropriate use of increasing hazard rate information, e.g., to decide when patients need special attention in clinical trials or to develop preventive maintenance programs. Increasing hazard rate information can be used to improve the effectiveness of any monitor that implicitly assumes a constant hazard, whether that monitor be a cumulative sum, an EWMA, or a more sophisticated Kalman / state space model. Section 4 derives a simultaneous EWMA for both mean and variance from a Bayesian sequential model.

Section 5 generalizes the EWMA to Kalman filtering. The development here parallels that of section 3, since an EWMA is a univariate Kalman filter. Section 5 illustrates how two alternative failure modes can be isolated using Kalman filtering. Section 6 applies the theory of section 5 to a more complex physical system, the air intake system for an automobile. Here, we find that the naive approach of adding extra parameters for alternative failure modes sometimes encounters numerical difficulties.

This occurs when the incoming observations do not provide sufficient information to simultaneously estimate all components of the enhanced state vector. In such cases, the natural alternative is to run multiple Kalman filters in parallel, each designed for a different combination of malfunctioning components; this is discussed in section 7. Concluding remarks appear in section 8.

This report is divided into three parts. The first part, sections 2-4, considers univariate monitors: Cusums and EWMAs. The second part, sections 5-7, discuss the special opportunities for detecting multiple faults for potentially univariate observations. The third very brief part presents concluding remarks.

We believe that the discussion in this report provides a strong argument for Bayesian sequential updating as a general, unifying principle for monitoring.

REFERENCES

Box, G., Graves, S., Bisgaard, S., Van Gilder, J., Marko, K, James, J., Seifer, M., Poublon, M., and Fodale, F. (2000) "Detecting Malfunctions in Dynamic Systems", *SAE Technical Paper Series* 2000-01-0363.

Box, G., Graves, S., Bisgaard, S., Graves, S., Kulahci, M., Marko, K, James, J., Van Gilder, J., Ting, T., Zatorski, H., and Wu, C. (2002) "Performance Evaluation of Dynamic Systems: The Waterfall Chart", *Quality Engineering* (to appear).

Girshick, M. A., and Rubin, H. (1952) "A Bayes Approach to a Quality Control Model", *Annals of Mathematical Statistics*, 23: 114-125.

Part I. Univariate Monitors

2. Bayes-Adjusted Cusum
3. Designing Bayesian EWMA Monitors Using Gage R & R and Reliability Data
4. Bayesian EWMA for Mean and Variance

We discuss here three univariate applications of Bayesian sequential updating. All assume univariate observations. The difference lies in the nature of the transition process. If the transition / deterioration occurs abruptly, Bayesian sequential updating produces a Bayes-adjusted Cusum, as described in section 2. If deterioration follows a normal random walk, Bayesian sequential updating gives us a Bayesian exponentially weighted moving average (EWMA), as we see in section 3. If in addition, the EWMA transition precision (reciprocal variance) itself migrates following a beta distribution, this principle generates a simultaneous EWMA for mean and variance, discussed in section 4. Part II generalizes the EWMA to a multivariate state space with possibly multivariate observations used for fault isolation.

2. BAYES-ADJUSTED CUSUM

Consider an abrupt jump from a simple (completely specified) good condition (H_0) to a simple bad condition (H_1). Bayesian sequential updating with independent, identically distributed (i.i.d.) observations and a constant hazard rate was considered by Girshick and Rubin (1952). They derived a rather simple but non-intuitive iteration from a cost model and a two-state, recurrent Markov chain; this was generalized to a continuous time stochastic process by Shiryaev (1963).

2.1. Non-i.i.d. observations, Non-Constant Hazard.

We wish to generalize the Girshick-Ruben procedure to non-i.i.d. observations and non-constant hazard. In particular, suppose we observe random variables y_t that have a density $f_{i,t} = f_{i,t}(y_t \mid y_{t-1}, y_{t-2},...)$, $i = 0, 1$ for good or bad (where $f_{0,t}$ and $f_{1,t}$ are densities with respect to a common dominating measure and may vary with t). Consider a random variable t_0 = the change point from good to bad. Let h_t = hazard rate = Pr{ bad at t +1 | good at t }. Further suppose that $f_{1,t}$ does not depend on t_0, which means that the Bayesian posterior can be summarized in a single number, g_t = Pr{ good at $t + 1 \mid y_1, ..., y_t$ }.

Step 1 of the two step iteration of Figure 1.1 requires us to compute the posterior probability of the "plant" being bad given the prior, g_{t-1}; we use the term "plant" to refer to the system being monitored, following the practice in the control theory literature. In this context, step 1, Bayes' theorem, gives us the following:

$$\Pr\{ \text{good at } t \mid y_1, ..., y_t \} = \frac{f_{0,t} g_{t-1}}{f_{0,t} g_{t-1} + f_{1,t}(1 - g_{t-1})}.$$

Step 2 allows for a possible transition, giving us the prior for the next observation as follows:

$$g_t = \Pr\{\text{ good at } (t+1) \mid y_1, \ldots, y_t \}$$

$$= (1 - h_t) \Pr\{\text{ good at } t \mid y_1, \ldots, y_t \} = \frac{(1-h_t) f_{0,t} g_{t-1}}{f_{0,t} g_{t-1} + f_{1,t}(1-g_{t-1})}. \qquad (2.1)$$

We now rewrite this in term of odds for the system being bad at time $(t + 1)$, $B_t = (1-g_t)/g_t$, as follows:

$$B_t = [\, h_t + (f_{1,t}/f_{0,t}) B_{t-1} \,] / (1 - h_t)$$

or

$$B_t = H_t + z_t B_{t-1}, \qquad (2.2)$$

where $H_t = h_t/(1 - h_t)$ = hazard odds, and $z_t = (f_{1,t}/f_{0,t})/(1 - h_t)$ = adjusted likelihood ratio. Without the transition (i.e., if $h_t = 0$), this is merely the odds formulation of Bayes' theorem: The posterior odds is the likelihood ratio times the prior odds.

Meanwhile, Girshick and Rubin converted (2.1) into a recursion for $Z_t = [1/g_t - 1/(1 - h_t)]/H_t = \{1 + B_t - [1/(1 - h_t)]\}/H_t$, which with non-constant hazard becomes

$$Z_t = z_t (1 + Z_{t-1})(H_{t-1}/H_t). \qquad (2.3)$$

Girshick and Rubin obtained the constant-hazard simplification of this: $Z_t = z_t(1 + Z_t)$.

Computationally, (2.2) and (2.3) often lead to numeric difficulties, which can be avoided by using logarithms. Let $\beta_t = \log(B_t)$ = log(odds for bad), $\eta_t = \log(H_t)$ = log(hazard odds), and $\zeta_t = \log(z_t)$ = log[likelihood ratio(t)] – log($1 - h_t$). Then (2.2) can be rewritten as

$$\beta_t = \eta_t + \log[1 + (z_t B_{t-1} / H_t)\,],$$

$$= \eta_t + \log[1 + \exp(\Delta_t)], \qquad (2.4)$$

where $\Delta_t = \zeta_t + \beta_{t-1} - \eta_t$. But ζ_t exceeds log[likelihood ratio(t)] by [– log(1 – h_t)] as long as $0 < h_t < 1$. However, in most practical applications, h_t is so small that $\log(1 - h_t) \cong (-h_t) \cong 0$, which makes ζ_t essentially the log(likelihood ratio).

An alternative to (2.4) can be obtained by factoring ($z_t B_{t-1}$) out of (2.2), producing the following:

$$\beta_t = \zeta_t + \beta_{t-1} + \log[1 + \exp(-\Delta_t)]. \tag{2.5}$$

We combine these two expressions using (2.4) when $0 > \Delta_t = (-|\Delta_t|)$ and using (2.5) when $0 \leq \Delta_t = |\Delta_t|$. This gives us

$$\beta_t = \max\{\, \eta_t,\, \zeta_t + \beta_{t-1} \,\} + \log[1 + \exp(-|\Delta_t|)]. \tag{2.6}$$

Note that for any Δ_t,

$$0 < \log[1 + \exp(-|\Delta_t|)] \leq \log(2);$$

except when $|\Delta_t|$ is small, this term will be negligible. In that case, (2.6) is a cumulative sum with a floor at η_t.

This can be written in a more familiar form by letting $Q_t^* = \beta_t - \eta_t$ = the excess over the log hazard odds of the log odds for the plant being bad. However, with non-constant hazard, β_t is still the log odds for the plant being bad, while Q_t^* no longer seems usable. Therefore, when writing Q_t^*, we shall henceforth assume constant hazard, $h_t = h$, so $Q_t^* = \beta_t - \eta$. Subtracting η from both sides of (2.6), we get

$$Q_t^* = \max\{\, 0,\, Q_{t-1}^* + \zeta_t \,\} + \log[1 + \exp(-|\Delta_t|)], \tag{2.7}$$

recalling from (2.4) that $\Delta_t = \zeta_t + \beta_{t-1} - \eta = Q_{t-1}^* + \zeta_t$. Therefore, this last expression becomes

$$Q_t^* = \max\{\, 0, \Delta_t \,\} + \log[1 + \exp(-|\Delta_t|)]. \tag{2.8}$$

Alternatively, we may write this as follows:

$$Q_t^* = \max\{\, 0,\ Q_{t-1}^* + \zeta_t \,\} + \log[1 + \exp(-|\, Q_{t-1}^* + \zeta_t \,|)]. \tag{2.9}$$

As mentioned above, the term $\log[1 + \exp(-|\, Q_{t-1}^* + \zeta_t \,|)]$ will be negligible except when $|\Delta_t|$ is small. If we drop it from (2.9), we get the following:

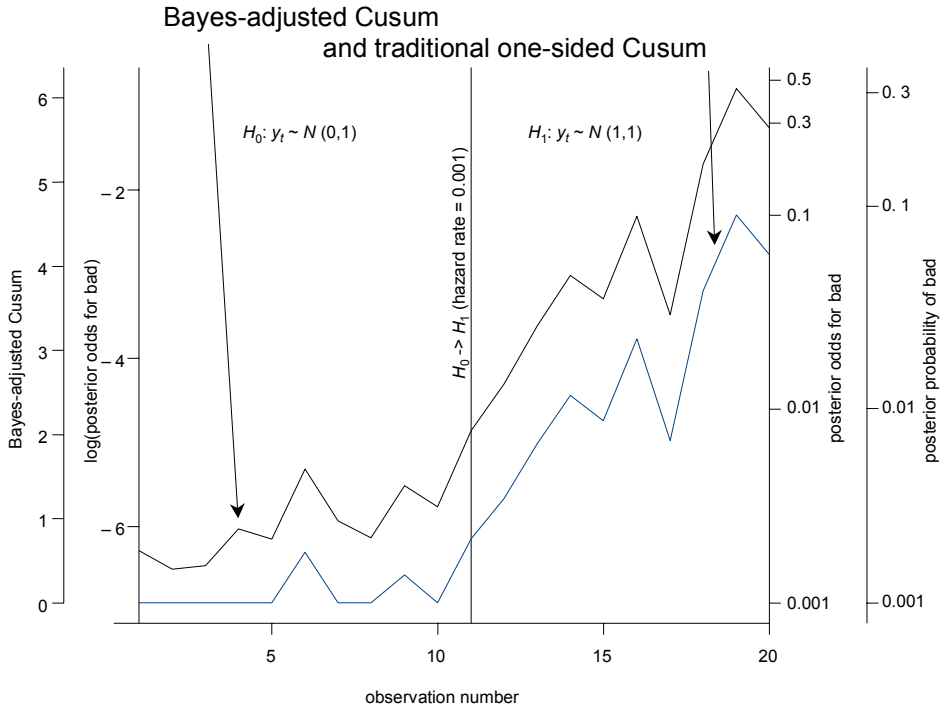
$$Q_t^+ = \max\{\, 0,\ Q_t^+ + \zeta_t \,\}. \tag{2.10}$$

Figure 2.1 presents a typical simulation comparing (2.9) and (2.10) with 20 observations changing from H_0 to H_1 at $t = 11$. In this, we assume $h_t = 0.001$, H_i: $y_t \sim N(\mu_i, 1)$, $i = 0, 1$, $\mu_0 = 0$, and $\mu_1 = 1$. First note that the log(likelihood ratio) in this case is given by the following:

$$\text{log(likelihood ratio)} = \left(y_t - \bar{\mu}\right) d \,/\, \sigma^2 , \tag{2.11}$$

where $\bar{\mu} = (\mu_1 + \mu_0)/2$ and $d = (\mu_1 - \mu_0)$. This makes Q_t^+ in Figure 2.1 a standard one-sided Cusum, apart from the term $[-\log(1 - h_t)]$ in ζ_t. However, since $h_t = 0.001$, we have $[-\log(1 - h_t)] = 0.001$, and this difference is not visually detectable in Figure 2.1.

Figure 2.1. One-Sided and a Bayes-Adjusted Cusum Simulations

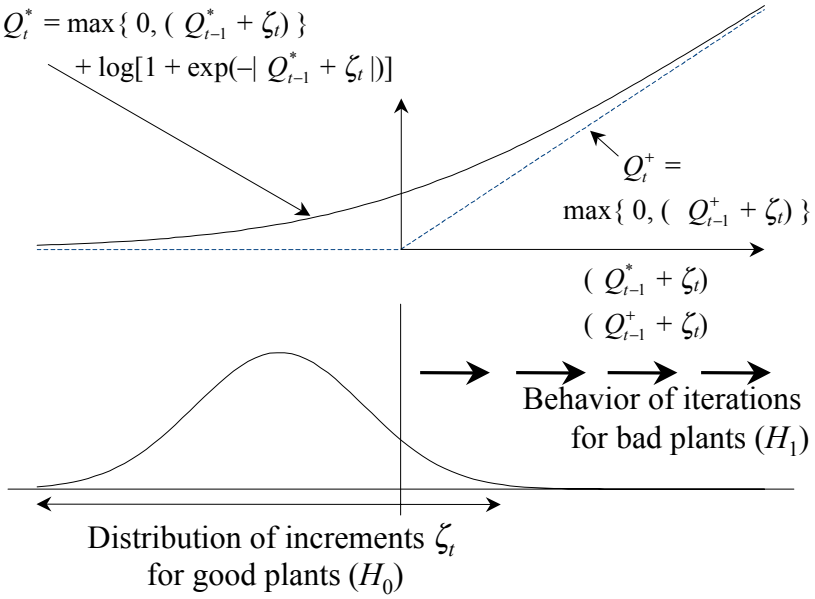


Four vertical scales are provided in Figure 2.1. The first is the "natural" Cusum scale, starting at 0. The second is the log odds for the plant being bad, which we obtain by solving for β_t the definition of the Bayes-adjusted Cusum with (2.7), getting $\beta_t = Q_t^* + \eta$, where $\eta = \ln[h/(1-h)] = \ln(0.001/0.999) = (-6.91)$. Recall that β_t is the log odds for the plant being bad at time $t + 1$, while Q_t^* is the excess in the log odds for bad over the log hazard odds. The third and fourth scales in Figure 2.1 simply translate the log odds for bad into odds, B_t, and probability, $(1 - g_t)$.

The behavior here is typical of what we have seen in other simulations with different values of μ_1 and differing numbers of observations before and after the change: Q_t^* and Q_t^+ tend to go up and down together. When the plant is bad, these movements are nearly identical. When the plant is good, the movements up and down of Q_t^* and Q_t^+

are not always parallel. Apart from the trimming effect of the max(.,.) function, this is due to the contribution of the Bayes-adjustment term $\log[1 + \exp(-|Q^*_{t-1} + \zeta_t|)]$ in (2.9). The effect of this term is displayed graphically in Figure 2.2.

Figure 2.2. Bayes-adjusted and Traditional Cusum Iteration



After the transition, under H_1, Q^+_t increases on average $E\zeta_t = d^2/(2\sigma^2) - \log(1 - h_t)$ each observation; meanwhile, Q^*_t increases slightly faster initially and approaches this average growth rate asymptotically; in the case considered in Figure 2.2, this asymptote was for practical purposes reached in one observation.

Other simulations suggest that the difference between the Bayes-adjusted and traditional Cusums seems large when the difference between good and bad is small and small when the difference between good and bad is large. However, even when the difference between good and bad is small to moderate, the difference between the Bayes-

adjusted and traditional Cusums seems to be fairly consistent and predictable, except when the threshold is low admitting a high false alarm rate.

This is consistent with previous reports, e.g., by Srivistava and Wu (1993, p. 665), who reported that with low thresholds and high false alarm rates, the Bayes' procedure was much better than the traditional Cusum. An earlier simulation comparison by Roberts (1966) found the Cusum and the Bayesian approaches essentially equivalent.

To be precise, Srivistava and Wu compared the "Shiryaev-Roberts" procedure to a Cusum and an exponentially weighted moving average (EWMA). Shiryaev (1963) assumed a uniform distribution for time to failure; this has a hazard rate $h(s) = 1/(t - s)$ that becomes infinite as $s \to t$. This was a reasonable theoretical contribution for 1963. However, more recent developments in biostatistics and reliability theory suggest that a hazard rate of this form would not be recommended except perhaps for some very special applications.

Fortunately, the Srivistava and Wu evaluation seems consistent with our intuition and with Roberts (1966), which suggests that these conclusions are relatively insensitive to the assumed uniform failure time distrbution. Roberts considered the Girshick-Rubin recursion (2.3), which is a linear transformation of the posterior odds. This transformation is constant only when the hazard rate is constant, which Girshick and Rubin assumed. However, with non-constant hazard, we are unable to find a sensible interpretation for either their iteration Z_t, (2.3), or our Bayes-adjusted Cusum Q_t^*, (2.7) - (2.8); in that case, we prefer the log odds for bad, β_t, using (2.6).

One advantage of the Girshick-Rubin iteration is that it produces an answer with zero hazard, which Roberts assumed. When the hazard rate is zero, we get different

answers from (2.5) and (2.7)-(2.9): Zero hazard means that the log hazard odds $\eta_t = (-\infty)$, so $\Delta_t = (+\infty)$, and (2.5) becomes

$$\beta_t = \zeta_t + \beta_{t-1}. \tag{2.12}$$

In words, the posterior log odds (β_t) is the prior log odds (β_{t-1}) plus the log likelihood ratio (ζ_t). This is the log odds formulation of Bayes' theorem and is also a traditional two-sided Cusum. It is clearly the correct answer if we are testing to evaluate an unchanging property of nature, as noted by Wald (1947).

However, any monitoring application involves a search for a change, which implies a non-zero hazard; in such situations, a traditional two-sided Cusum, (2.12), would be inappropriate except when used with a traditional V-mask, which makes it equivalent to two one-sided Cusums.

For zero hazard, both the Girshick-Rubin iteration (2.3) and our Bayes-adjusted Cusum Q_t^*, (2.7) - (2.9), involve indeterminate forms: $(H_{t-1}/H_t) = (0/0)$ in (2.3) and ($\eta_t - \eta_{t-1}$) = [(– ∞) – (– ∞)] in (2.9). We can avoid these indeterminate forms by assuming the hazard rate is non-zero and constant but virtually negligible. For constant hazard, our Bayes-adjusted Cusum, Q_t^*, is a monotonic transformation of the Girshick-Rubin iteration, and with small hazard, both are essentially equivalent to the standard one-sided Cusum (2.10), introduced by Page (1954).

In sum, if the hazard rate is truly zero, then a two-sided Cusum performing a Bayesian formulation of a Wald sequential test (2.12) is appropriate. Meanwhile a non-zero but constant and small hazard rate calls for a monitor that is virtually equivalent to a one-sided Cusum (2.10).

In this subsection, we discussed a Bayesian iteration for an abrupt jump from good to bad and displayed similarities to a Cusum. We next consider selection and interpretation of a detection threshold.

2.2. Cusum Threshold and Increase in Posterior Log(Odds)

An obvious decision criterion for Bayesian monitoring is to set an alarm when the posterior probability of the plant being bad exceeds a threshold, which may be tied to the economics of the problem (e.g., Girshick and Rubin 1952). This translates into a threshold for the posterior log odds for bad, β_t. With constant hazard, this is equivalent to setting a threshold for our Bayes-adjusted Cusum, $Q_t^* = \beta_t - \eta$; this threshold on Q_t^* becomes the "increase" in log odds for a bad plant (β_t) over the log hazard odds (η) required to set an alarm.

The substantial similarities between a one-sided Cusum Q_t^+ and the Bayes monitor Q_t^* suggest that the detection threshold for the traditional Cusum Q_t^+ is roughly equivalent to that for Q_t^*, although the example of Figure 2.2 and similar simulations indicates a consistent bias between Q_t^* and Q_t^+. Other simulations like Figure 2.2 suggest that this bias may decline as the difference between H_0 and H_1 increases. Further research is needed to determine the magnitude of this bias and the extent to which this odds-increase interpretation can be extended from Q_t^* to Q_t^+.

Ignoring this bias, this equivalence is illustrated in Table 2.1 for a constant hazard rate of 0.01. To deepen our understanding of this connection, suppose we are collecting

one sample per day from a sewage treatment plant and preparing a Cusum chart of the result. And suppose that $\zeta_t = \log(f_{1,t}/f_{0,t}) - \log(1 - h_t)$ in (2.9) has standard deviation of 1 and mean (– 0.5) if the system is good and (+0.5) if bad. A standard Cusum chart might put the threshold for Q_t^+ at 4, which would have an Average Run Length (ARL) of roughly 350 to a false alarm and 8.5 to a valid alarm (e.g., Bissell 1969).

Table 2.1. Equivalent Thresholds between a One-Sided Cusum and the Bayes' Posterior

One-Sided	Bayes Posterior with Hazard 0.01		
Cusum	log(odds)	odds	probability
3	– 1.60	0.20	0.17
4	– 0.60	0.55	0.36
5	0.40	1.50	0.60

(Caveat: These numbers are intended only to illustrate the essential equivalence of the different interpretations. Further research is needed to model the bias visible in Figure 2.1 but ignored here.)

Suppose also that the plant has an upset (goes bad) roughly once every 100 days, with roughly a constant hazard rate of 0.01. Then from Table 2.1, we see that this is roughly equivalent to deciding to declare an upset when the posterior probability of the system being bad is 0.36 or greater. Note that if the model is correct, then this 0.36 reflects the proportion of systems with comparable histories that are bad; it is not (merely) a subjective probability. (As noted with the table, further research is needed to adequately model the bias visible in Figure 2.1 but ignored in Table 2.1.)

2.3. Costs and Run Lengths

In many applications, the expected cost of a delay to detection will be proportional to the average run length (ARL) to an alarm after the system monitored becomes bad. Meanwhile, the expected cost of a false alarm might be proportional to the probability of an alarm during the good life of the system monitored (the "plant"). Obviously, increasing the threshold increases the ARL(bad) while reducing the false alarm rate. Therefore, selecting a threshold implies a certain assessment of the cost of a false alarm relative to the cost of a delay of one more observation. This gives us three equivalent ways to select a threshold for a Cusum / Bayes monitor:

(*a*) Select a posterior probability [or log(odds)] above which an alarm is declared.

(*b*) Select a threshold to balance some characteristics of the run length distributions for good and bad systems.

(*c*) Specify the cost of a false alarm relative to the cost of waiting one more observation before declaring an alarm on a bad system.

In practical monitor design, it may be wise to evaluate all three perspectives before making the final choice of threshold.

This three part equivalence assumes the model is correct, which often is not the case. For example, a monitor may be designed ignoring serial dependencies in the data, because it is not feasible to model them with the resources available in the project. The final evaluation and selection of a threshold might be made by extrapolating from run length data collected using artificially low thresholds in prototype systems, as suggested by Bisgaard et al. (2001). These thresholds would automatically adjust for serial

dependence and model inadequacy in the data. Meanwhile, the threshold implied by this equivalence for the posterior probability of the system being bad, ignoring serial dependence and model inadequacy, might be ridiculously close to 1; in such cases, the formally computed posterior is not a realistic assessment of the relative frequency of systems with comparable histories that are bad. This does not negate the value of the monitor, only the posterior probability interpretation of it.

2.4. Cuscores

We now return to the question of monitor design. In some applications, it is more convenient to think in terms of looking for a signal in noise than to think in terms of defining "bad" distinct from "good". In such cases we replace the log(likelihood ratio) in (2.10) with Fisher's efficient score [ignoring the term $\log(1 - h_t)$ in ζ_t]. This gives us

$$Q_t^+ = \max\left\{0, Q_{t-1}^+ + \left.\frac{\partial \log\left[f_t\left(y_t \mid \theta\right)\right]}{\partial \theta}\right|_{\theta=0}\right\}, \tag{2.13}$$

where θ is a parameter indicating the extent to which a malfunction or signal of interest is present; this is called the cumulative score function or Cuscore by Box and Ramírez (1992) and Box and Luceño (1997). In essence, (2.13) is tangent to the log(likelihood), while (2.10) is a secant line. The choice between the two rests more on which form seems to relate most easily to the way a user thinks about a particular monitoring situation, with neither form being universally preferred. For further discussion of the use of these techniques for detecting changes in regression or time series models, see Box and Ramírez (1992) or Box and Luceño (1997).

We note in passing that the term "Cuscore" has been used in a different sense by others, e.g., Radaelli (1992), to denote a Cusum of "scores" achieved by transforming a continuous variable into a small number of discrete categories, roughly rounding off the observations grossly to 0 and 1. This has the advantage of simplifying the analysis of the run length distribution. However, it throws away some portion of the information in the observations in order to do so.

2.5. If the Bad Distribution Depends on the Changepoint

Finally, we recall that the discussion so far ruled out one class of abrupt jumps from a simple good H_0 to a simple bad H_1: situations such as tool wear, for which $f_{1,t}$ may depend on the changepoint t_0, and for which the posterior can not be summarized in one number. In such cases, we suggest that Bayesian sequential updating still provides a theoretical best procedure. This theoretical best is not computationally feasible but still has value for evaluating the relative responsiveness of computational procedures.

2.6. Discussion

In this section, we looked for an abrupt jump from one simple model of good to a simple model of bad. We found that when the hazard rate is low and relatively constant, Bayesian sequential updating is reasonably well approximated by a Cusum of log(likelihood ratio). If the hazard rate is not constant (i.e., the distribution of time to a problem is not exponential), then the theory presented here provides a natural way for improving diagnostic performance through the use of that information. This could be

quite valuable for designing preventive maintenance procedures that combine reliability models of equipment that wears out or ages in other ways with periodic data collection. This holds promise for developing procedures that outperform any procedure that considers only one source of information.

The next section of this report considers a random walk as a model for systems that tend to fail through random, gradual deterioration instead of an abrupt jump.

REFERENCES

Bisgaard, S., Graves, S., Kulahci, M., Van Gilder, J., James, J., Marko, K., Ting, T., Wu, C., and Zatorski, H. (2001) "Accelerated Testing of On-Board Diagnostics" (unpublished manuscript).

Bissell, A. F. (1969) "Cusum Techniques for Quality Control", *Applied Statistics*, 18: 1-30.

Box, G., and Luceño, A. (1997) *Statistical Control by Monitoring and Feedback Adjustment* (NY: Wiley).

Box, G., and Ramirez, J. (1992) "Cumulative Score Charts", *Quality and Reliability International*, 8, 17-27.

Girshick, M. A., and Rubin, H. (1952) "A Bayes Approach to a Quality Control Model", *Annals of Mathematical Statistics*, 23: 114-125.

Page, E. S. (1954) "Continuous Inspection Schemes", *Biometrika*, 41, 100-115.

Radaelli, G. (1992) "Using the Cuscore Technique in the Surveillance of Rare Health Events", *Journal of Applied Statistics*, 19, 75-81.

Roberts, R. W. (1966) "A Comparison of Some Control Chart Procedures", *Technometrics*, 8: 411-430.

Shiryaev, A. N. (1963) "On Optimal Methods in Quickest Detection Problems", *Theory of Probability and Its Applications*, 8, 22-46.

Srivistava, M. S., and Wu, Y. (1993) "Comparison of EWMA Cusum, and Shyryayev-Roberts Procedures for Detecting a Shift in the Mean", *Annals of Statistics*, 21: 645-670.

Wald, A. (1947) *Sequential Analysis* (NY: Wiley; reprinted 1973 by Dover).

3. DESIGNING BAYESIAN EWMA MONITORS USING GAGE R & R AND RELIABILITY DATA

Consider a physical system with a condition x_t that is not directly observable but that is assumed to follow a random walk, as

$$x_t = \mu_t + x_{t-1} + w_t, \quad w_t \sim N(0, \sigma^2_{w,t}), \tag{3.1}$$

where μ_t represents a potential deterministic drift, and w_t represents an unpredictable portion of system reliability. In subsection 3.1, we show how μ_t and $\sigma_{w,t}$ relate to the reliability hazard rate. In particular, we show that any reliability distribution can be modeled in terms of $\sigma_{w,t}$ or $(\mu_t, \sigma_{w,t})$.

The analyses of this report assume we can model adequately the distribution of x_t at first use, $t = 1$. No monitor is designed for a completely unprecedented application: The time and money to design and use a monitor is justified from experience with other applications, which can be used to estimate a distribution at first use. For a manufactured product, this could be obtained from control chart data collected at the end of the production line. If this is a new product, the distribution at first use could be estimated from the history of similar products, adjusted if appropriate considering design objectives, prototype test data, and previous new product introductions. In biostatistics, it could be obtained from previous clinical trials of roughly comparable therapies. In portfolio management, one could consider the behavior of similar financial instruments. In this article, we shall assume that $x_1 \sim N(x_{1|0}, \sigma^2_{1|0})$.

In subsection 3.2, we consider a process of the form (3.1) (with $\mu_t = 0$ and $\sigma_{w,t} = \sigma_w$ constant) observed with error,

$$y_t = x_t + v_t, \quad v_t \sim N(0, \sigma_v^2). \tag{3.2}$$

In many situations, σ_v can be estimated from a study of gage repeatability and reproducibility (NIST 2001, ch. 2).

With adequate estimates of the distribution at first use, the hazard rate and σ_v, we can estimate the relative frequency distribution of the condition of plants at any future point in time, among all plants with comparable observed histories. Subjective probabilities can be used, but objective probabilities are also available, within the limits of estimation precision and the comparability of our initial reference set.

The conceptual framework is outlined in Figure 3.1, with mathematical details summarized in Table 3.1. The result is an exponentially weighted moving average (EWMA), except that the weight on the last observation varies with time, converging to a constant; expression numbers in Table 3.1 are keyed to the discussion below.

Figure 3.1. Bayesian Sequential Updating with a Random Walk

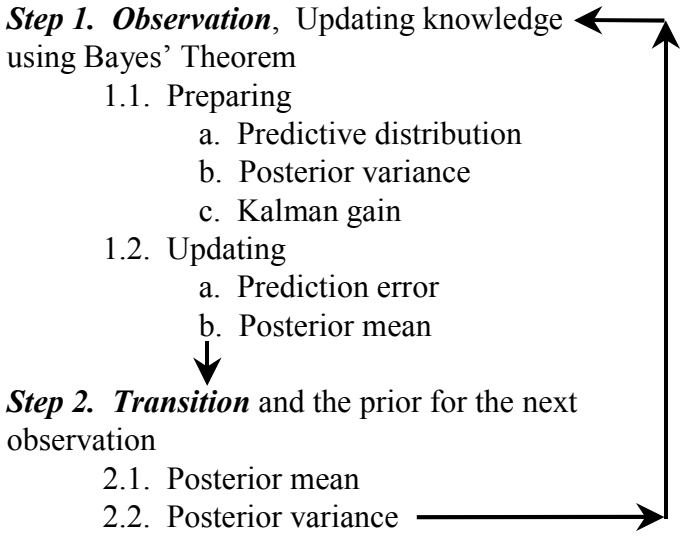


Table 3.1. Bayesian EWMA Computations

Prediction $x_{t+1|t}$ for time $(t + 1)$ given information $D_t = \{y_t, y_{t-1}, ...\}$, available at time t:

$$x_{t+1|t} = (1 - K_t)x_{t|t-1} + K_t y_t = x_{t|t-1} + K_t e_t, \;\; e_t = (y_t - x_{t|t-1}) \qquad (3.18)$$

Weight on the last observation (Kalman gain):

$$K_t = 1 \,/\, \{\, 1 + [1/(\rho^2 + K_{t-1})] \,\} \qquad (3.20)$$

where

$\rho^2 = \sigma_w^2 / \sigma_v^2$ = (migration variance) / (measurement variance)

and

$K_1 = 1$ with no prior knowledge of the initial condition of the plant

Confidence bounds on x_t are obtained from $(x_t \mid D_{t-1}) \sim N(x_{t|t-1},\ \sigma^2_{t|t-1})$, where

$$\sigma^2_{t+1|t} = \left(\sigma^{-2}_{t|t-1} + \sigma_v^{-2}\right)^{-1} + \sigma_w^2 \qquad (3.11) \;\&\; (3.17)$$

Evaluate prediction error e_t relative to

$$\text{var}(y_t \mid D_{t-1}) = \sigma^2_{t|t-1} + \sigma_v^2 \qquad (3.9)$$

This procedure provides a fast initial response (FIR) approach that is different from the FIR technique in the literature and is clearly tied to information obtainable for virtually any monitoring application: the distribution of the condition of the plant at first use, x_1, the reliability, coded in σ_w, and the measurement noise σ_v. If the condition at first use is non-informative, this theory sets $K_1 = 1$ and has K_t declining monotonically to an asymptote as information about x_t accumulates. We believe this has much greater intuitive appeal than the traditional FIR approach, which essentially asserts that the prior at $t = 1$ is as informative as the prior at any $t > 1$.

This FIR approach is discussed in section 3.5 after considering an example in section 3.3 and discussing further the asymptotic behavior of the algorithm in section 3.4.

Questions of robustness are considered in section 3.6. Section 3.7 considers when to declare a malfunction, and a summary discussion of this "normal random walk observed with error" appears in section 3.8. This procedure is generalized to estimate a changing variance as well as a mean in section 4 and to sequential reestimation of gradually changing regression parameters in sections 5-7; these later sections consider fault isolation as well as detection. Much of this material is also known as Kalman filtering, although our Bayesian development differs somewhat from the minimum mean square prediction error principle used by Kalman (1960).

3.1. Hazard and Migration Rates

We show here how the migration parameters (μ_t, $\sigma_{w,t}$) in (3.1) determine the reliability distribution, expressed in the hazard rate h_t, and how the hazard rate constrains μ_t and determines $\sigma_{w,t}$ given μ_t. Of course, complex systems, whether products, production processes or humans, can be impacted by many different kinds of problems. We assume that h_t is the hazard rate relevant to a process x_t observed indirectly via y_t. Standard techniques in biostatistics and reliability support cause-specific estimation of hazard rates.

To understand the relationship between (μ_t, $\sigma_{w,t}$) and h_t, we start by assuming that x_t is good as long as $L \le x_t \le U$. If the distribution at first use is $x_1 \sim N(x_{1|0}, \sigma^2_{1|0})$, then $h_1 = h_{0,1} + h_{1,1}$, where

$$h_{0,1} = \Phi\left(\frac{L - x_{1|0}}{\sigma_{1|0}}\right),$$

and

$$h_{1,1} = \left[1 - \Phi\left(\frac{U - x_{1|0}}{\sigma_{1|0}}\right)\right]. \tag{3.3}$$

For a manufactured product, h_1 is the proportion of units with x_1 outside (L, U) or that fail for this reason when the customer first attempts to use them.

Let $F_t(x_t) = F_t(x_t \mid L \le x_\tau \le U, \text{ for all } \tau \le t)$ be the cumulative distribution function (cdf) for x_t given that it is not bad and has not previously been bad. Then

$$F_1(x_1) = \begin{cases} 0 & \text{if } x_1 < L \\ \dfrac{\Phi\left(\dfrac{x_1 - x_{1|0}}{\sigma_{1|0}}\right) - h_{0,1}}{1 - h_1} & \text{if } L \le x_1 < U \\ 1 & \text{if } U \le x_1 \end{cases} \tag{3.4}$$

Starting from (3.3) and (3.4), we derive the hazard rate h_t and the cdf for x_t good $F_t(x_t)$ recursive in pieces as follows. First, let $F_{t,0}(x_t)$ be the cdf for x_t, good or bad at time t given that it has not previously been bad as

$$F_{t,0}(x_t) = \int \Phi\left(\frac{x_t - x_{t-1} - \mu_t}{\sigma_{w,t}}\right) dF_{t-1}(x_{t-1}) . \tag{3.5}$$

Then the proportions of units too small and too large at time t among those good at $t - 1$ are

$$h_{0,t} = F_{t,0}(L),$$

and

$$h_{1,t} = 1 - F_{t,0}(U).$$

The hazard rate at time t is the sum of those failing both small and large, as

$$h_t = h_{0,t} + h_{1,t}, \tag{3.6}$$

and the distribution of those still good is the truncated distribution from $F_{t,0}$, as

$$F_t(x_t) = \begin{cases} 0 & \text{if } x_t < L \\ \left[\dfrac{F_{t,0}(x_t) - h_{0,t}}{1 - h_t} \right] & \text{if } L \le x_t < U \\ 1 & \text{if } U \le x_t \end{cases}. \tag{3.7}$$

From (3.3) - (3.7), we see that the sequence (μ_t, $\sigma_{w,t}$) uniquely determines the hazard rate. Conversely, if $L \le x_{1|0} < U$ and $\mu_t = 0$ for all t, then h_t is monotonically increasing in $\sigma_{w,t}$ and is 0 when $\sigma_{w,t}$, = 0, which means that h_t uniquely determines $\sigma_{w,t}$. [Expressions (3.1) and (3.3) - (3.7) can be generalized to a multivariate state space by assuming that $\mathbf{x}_t$ is bad if it is outside an acceptance region A and defining F_t in the obvious way for all Borel sets. In this more general setting, the parameters of the transition distributions uniquely determine the hazard rate. With suitable additional restrictions, the hazard rate can uniquely determine some univariate aspect of the migration distribution.]

Therefore, given data on product reliability or time to onset of an adverse reaction in clinical trials, a reasonably parsimonious model can be built for (μ_t, $\sigma_{w,t}$) consistent with the available data. This does not require data on a new product or therapy never used before; it only requires data on comparable products or therapies currently in use.

3.2. Univariate Bayesian Updating and an EWMA

Now suppose we have y_t being a noisy observation of the unknowable state of the plant x_t, per (3.2). We shall apply Bayesian sequential updating to this example with the added simplifications of assuming $\mu_t = 0$ and $\sigma_{w,t} = \sigma_w$ = constant. We shall find that this gives us an exponentially weighted moving average (EWMA) in the limit for large t with an intuitively satisfying Bayesian answer to the fast initial response (FIR) problem. This

approach will be compared to the traditional FIR in section 3.4 below, after discussing threshold selection in section 3.3.

For cases involving normally distributed noise in observation and transition, we find it convenient to break the two steps in Figure 1.1 into the substeps outlined in Figure 3.1. In particular, we divide step 1 into "1.1. Preparing" and "1.2. Updating". This distinction highlights the fact that "1.1. Preparing" can take place between the previous execution of step 2 and the current "1.2. Updating" step. With stationary systems most of "1.1. Preparing" can be computed offline in advance of the application. With traditional Kalman filtering (and traditional EWMAs), the Kalman gain is often replaced by an asymptotic value, and its transients are ignored. This can help reduce demands on a real-time microprocessor, allowing in some cases the use of a cheaper microprocessor.

We will preface this development with a brief comment about notation: As indicated with (3.1) and (3.2), observations y_t provide information about an unknown state of nature x_t. Just before each observation arrives, our knowledge of x_t is summarized in the prior $(x_t \mid D_{t-1}) \sim N(x_{t|t-1},\ \sigma^2_{t|t-1})$, where $D_{t-1} = \{y_{t-1}, y_{t-2}, \ldots y_1, x_{1|0}, \sigma^2_{1|0}\}$; at time $t = 1$, this is the distribution at first use. Step 1 in Figure 3.1 transforms this prior into the posterior $(x_t \mid D_t) \sim N(x_{t|t},\ \sigma^2_{t|t})$. Step 2 then models a transition from x_t to x_{t+1}, and our knowledge then degrades accordingly to $(x_{t+1} \mid D_t) \sim N(x_{t+1|t},\ \sigma^2_{t+1|t})$, which becomes the prior at the next point in time. We now consider specifics of these steps.

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

Step 1.1a. Predictive Distribution. We begin by combining $(x_t \mid D_{t-1}) \sim N(x_{t|t-1}, \sigma^2_{t|t-1})$ with the observation process (3.2) and integrating out the unknowable x_t to get the predictive distribution as follows:

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

where

$$f_t = x_{t|t-1}, \tag{3.8}$$

since the expected value of the sum in (3.2) is the sum of the expectations, and

$$\sigma^2_{y|t-1} = \sigma^2_{t|t-1} + \sigma^2_v, \tag{3.9}$$

since the variance of a sum of uncorrelated random variables is the sum of the variances. (We maintain the distinction between f_t and $x_{t|t-1}$ to stress their different functions and to facilitate generalization to situations where they are different.)

Step 1.1b. Posterior Variance. When a quantity with a normal prior is observed with additive normal error, the posterior is also normal. The posterior mean is a weighted average of the prior mean and the observation with weights inversely proportional to the variances. The posterior variance is the reciprocal sum of the reciprocal variances (e.g., DeGroot 1970, p. 167). We shall write this as follows:

$$(x_t \mid D_t) \sim N(x_{t|t}, \sigma^2_{t|t}),$$

where

$$x_{t|t} = \frac{\sigma^{-2}_{t|t-1} x_{t|t-1} + \sigma^{-2}_v y_t}{\sigma^{-2}_{t|t-1} + \sigma^{-2}_v} = \sigma^2_{t|t} \left\{ \sigma^{-2}_{t|t-1} x_{t|t-1} + \sigma^{-2}_v y_t \right\}, \tag{3.10}$$

and

$$\sigma^{-2}_{t|t} = \sigma^{-2}_{t|t-1} + \sigma^{-2}_v. \tag{3.11}$$

DeGroot (1970, p. 38) calls a squared reciprocal scale factor a "precision", which for the normal distribution is one over the variance. With this concept, (3.11) says that

the posterior precision, $\sigma_{t|t}^{-2}$, is the sum of the precisions of the prior, $\sigma_{t|t-1}^{-2}$, and the observation, σ_v^{-2}.

Step 1.1c. Kalman Gain. The weight on the last observation y_t in (3.10) is called the Kalman gain, and will be denoted as follows:

$$K_t = \sigma_{t|t}^2 \sigma_v^{-2}. \tag{3.12}$$

From (3.11), we see that

$$\sigma_{t|t-1}^{-2} = \sigma_{t|t}^{-2} - \sigma_v^{-2}.$$

We substitute these last two expressions into (3.10) to get

$$\begin{aligned} x_{t|t} &= \sigma_{t|t}^2 \left\{ \left(\sigma_{t|t}^{-2} - \sigma_v^{-2} \right) x_{t|t-1} + \sigma_v^{-2} y_t \right\} \\ &= x_{t|t-1} + K_t \left(y_t - x_{t|t-1} \right). \end{aligned} \tag{3.13}$$

For plants with stationary transitions and constant observation and transition variances, all the computations of substep 1.1 can be done offline except for the mean of the predictive distribution. 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 compute the prediction error and use that to update the "posterior mean", our point estimate of the state of the plant.

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

$$e_t = y_t - f_t. \tag{3.14}$$

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 (3.13), as

$$x_{t|t} = x_{t|t-1} + K_t e_t . \tag{3.15}$$

This completes step 1, observation, in Bayesian sequential updating as outlined in Figure 3.1. Next, we permit the plant to transition in preparation for the next observation, per step 2.

Step 2. Transition and Prior for the Next Observation. Given the posterior mean and variance from step 1, we can easily compute using (3.1) the prior mean and variance for the next observations, as follows:

Step 2.1. Prior Mean.

$$x_{t+1|t} = x_{t|t}, \tag{3.16}$$

and

Step 2.2. Prior Variance.

$$\sigma^2_{t+1|t} = \sigma^2_{t|t} + \sigma^2_w . \tag{3.17}$$

This completes step 2. The resulting prior distribution at one point in time $N(x_{t+1|t}, \sigma^2_{t+1|t})$ becomes an input for step 1.1, $N(x_{t|t-1}, \sigma^2_{t|t-1})$, at the next point in time. In this way, observations are processed sequentially as they arrive. If the model (3.1) - (3.2) is correct, then the prior $N(x_{t+1|t}, \sigma^2_{t+1|t})$ summarizes all the information in $D_t = \{y_t, y_{t-1}, ..., y_1, x_{1|0}, \sigma_{1|0}\}$ about the state of the plant at time t+1.

The most important expressions in this section are summarized in Figure 3.2. We next apply this iteration to an example (section 3.3) before deriving some general properties of this case. These properties include the convergence of the Kalman gain to an asymptote (section 3.4). This convergence turns out to be monotonic, which helps to establish it as a natural Bayesian answer to the fast initial response (FIR) problem. This is followed by discussions of robustness (section 3.6), threshold selection (section 3.7), and concluding remarks on EWMAs (section 3.8).

Figure 3.2. Bayesian EWMA Iteration

Step 1. Observation, updating knowledge using Bayes' theorem

1.0. Observation model

a. Prior $(x_t \mid D_{t-1}) \sim N(x_{t|t-1},\ \sigma^2_{t|t-1})$ (from step 2)

b. Observation $(y_t \mid x_t) \sim N(x_t,\ \sigma^2_v)$ (3.2)

1.1. Preparing

a. Predictive distribution $(y_t \mid D_{t-1}) \sim N(f_t,\ \sigma^2_{y|t-1})$,

$$f_t = x_{t|t-1},\ \sigma^2_{y|t-1} = \sigma^2_{t|t-1} + \sigma^2_v \quad (3.9)$$

b. Posterior precision and variance $\sigma^{-2}_{t|t} = \sigma^{-2}_{t|t-1} + \sigma^{-2}_v$ (3.11)

c. Kalman gain $K_t = \sigma^2_{t|t}\sigma^{-2}_v$ (3.12)

1.2. Updating

a. Prediction error $e_t = y_t - f_t$ (3.14)

b. Posterior mean $x_{t|t} = x_{t|t-1} + K_t e_t$ (3.15)

Step 2. Transition and the prior for the next observation

2.0. Model $(x_{t+1} \mid x_t) \sim N(x_t,\ \sigma^2_w)$ (3.1)

2.1. Posterior mean $x_{t+1|t} = x_{t|t}$ (3.16)

2.2. Posterior variance $\sigma^2_{t+1|t} = \sigma^2_{t|t} + \sigma^2_w$ (3.17)

In sum: Combining (3.13) - (3.15):

$$x_{t+1|t} = x_{t|t-1} + K_t\left(y_t - x_{t|t-1}\right) = \left(1 - K_t\right)x_{t|t-1} + K_t y_t \quad (3.18)$$

where

$$K_t = 1\Big/\left\{1 + \left[1\Big/\left(\rho^2 + K_{t-1}\right)\right]\right\},\ \rho = \sigma_w/\sigma_v \quad (3.20)$$

$$\rightarrow K_\infty = \left(\rho^2/2\right)\left\{\sqrt{1 + \left(4/\rho^2\right)} - 1\right\} \quad (3.21)$$

For confidence limits, combine (3.11) and (3.17):

$$\sigma^2_{t+1|t} = \left(\sigma^{-2}_{t|t-1} + \sigma^{-2}_v\right)^{-1} + \sigma^2_w$$

3.3. Sample Computations for a Bayesian EWMA

Sample computations using this procedure are given in Table 3.2 and Figure 3.3. As suggested in the summary box at the bottom of Figure 3.2, we really only need three columns from Table 3.2.2: the prior mean and variance and the Kalman gain. The remaining columns of Table 3.2.2 are provided to describe more clearly the machinery of Bayesian updating as discussed with Figures 3.2 and 3.1.

Table 3.2. Univariate Bayesian Sequential Updating: Illustrative Calculations

Table 3.2.1. Scenario Simulated

Manufacturing distribution

	Mean	Variance	Standard Deviation	Precision
	$x_{1\|0}$	$\sigma^2_{1\|0}$	$\sigma_{1\|0}$	$\sigma^{-2}_{1\|0}$
	0	0.1	0.316	10

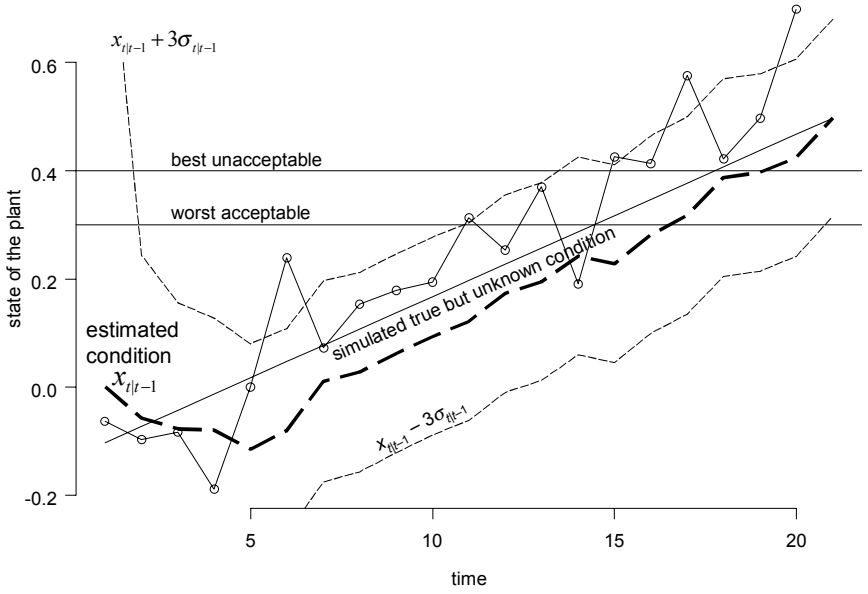
Observation error	Variance	Standard Deviation	Precision
	σ^2_v	σ_v	σ^{-2}_v
	0.01	0.1	100

Migration	Mean	Variance	Standard Deviation	Precision
	μ_t	σ^2_w	σ_w	σ^{-2}_w
"Actual"	0.03	0	0	∞
Assumed	0	0.001	0.0316	1,000

Table 3.2.2. Illustrative Calculations

	Simulated		Prior from Previous Step 2			Intermediate Computations in Step 1			
						Posterior		Kalman	Prediction
Time	True State x_t	Observation $y_t = x_t + v_t$	Mean $x_{t\|t-1}$	Variance $\sigma^2_{t\|t-1}$	Precision $\sigma^{-2}_{t\|t-1}$	Precision $\sigma^{-2}_{t\|t}$	Variance $\sigma^2_{t\|t}$	Gain K_t	Error e_t
eq'n→	(3.1)	(3.2)	(3.16) [+(3.15)]	(3.17)		(3.11)		(3.12)	(3.14)+ (3.8)
1	– 0.103	– 0.063	0	0.1000	10.0	110.0	0.00909	0.909	– 0.063
2	– 0.073	– 0.097	– 0.057	0.0101	99.1	199.1	0.00502	0.502	– 0.040
3	– 0.043	– 0.084	– 0.077	0.0060	166.0	266.0	0.00376	0.376	– 0.007
...									...
19	0.437	0.497	0.396	0.0037	270.2	370.2	0.00270	0.270	0.101
20	0.467	0.698	0.423	0.0037	270.2	370.2	0.00270	0.270	0.275
21	0.497		0.497	0.0037	270.2				

Figure 3.3. A Bayesian EWMA



Numbers for simulated "true state" and "observation" are given in the second and third columns of Table 3.2.2. These were obtained as pseudo-random numbers generated according to the "manufacturing distribution", "observation error", and "migration" described in Table 3.2.1. The manufacturing distribution is assumed to be normal with mean, variance, and precision as given in Table 3.2.1; these define the prior at time $t = 1$. We begin computing the "posterior precision" for observation 1 using expression (3.11), as $\sigma_{1|1}^{-2} = \sigma_{1|0}^{-2} + \sigma_v^{-2} = 10 + 100 = 110$. The posterior variance $\sigma_{1|1}^2 = 1/110 = 0.00909$. The Kalman gain is obtained from (3.12) as $K_t = \sigma_{1|1}^2 \, \sigma_v^{-2} = 0.00909 \times 100 = 0.909$. The prediction error (–0.063) is obtained as usual as the observation (–0.063) minus the forecast 0, per (3.14) and (3.8). The posterior mean, per (3.15), is then the prior mean plus $K_t \, e_t = 0 + (0.909) \times (–0.063) = (–0.057)$; this appears in Table 3.2.2 as the prior for time $t = 2$, per (3.16).

Note that the prior and posterior variances and precisions, $\sigma^2_{t|t-1}$, $\sigma^2_{t|t}$, $\sigma^{-2}_{t|t-1}$, and $\sigma^{-2}_{t|t}$, and the Kalman gain, K_t, all converge to constants to three significant digits by observation $t = 20$. This occurs here because σ_v and σ_w are constant. This is a special case of a more general result that for "completely observable" models (Gelb 1999, p. 142) with constant, linear transitions and constant observation and transition covariance matrices, the Kalman gain and the prior and posterior covariance matrices all converge to constants. We next consider more carefully the behavior of K_t in this EWMA case.

3.4. Kalman Gain for a Bayesian EWMA

In this section, we study the behavior over time of the Kalman gain of (3.12) and tie more carefully the above model to a traditional exponentially weghted moving average (EWMA). First, we combine (3.13) with (3.16) to obtain the following:

$$\begin{aligned} x_{t+1|t} &= x_{t|t-1} + K_t(y_t - x_{t|t-1}) \\ &= (1 - K_t)x_{t|t-1} + K_t y_t \,. \end{aligned} \tag{3.18}$$

This shows more clearly than (3.10) that the posterior mean $x_{t+1|t}$ is a weighted average of $x_{t|t-1}$ and y_t. By recursively substituting (3.18) into itself, we can show that $x_{t+1|t}$ is a weighted average of y_{t-j}, $j = 0, 1, \ldots$, with weights declining exponentially, provided $0 < K_t < 1$ and K_t is bounded away from 1 as $t \to \infty$. [Box and Luceño (1997, p. 69, 91) discuss this assuming $K_t = K_\infty$ is constant.]

To confirm that $0 < K_t < 1$, substitute (3.11) into (3.12) to obtain the following:

$$K_t = \sigma^2_{t|t}\sigma^{-2}_v = \left(\sigma^{-2}_{t|t-1} + \sigma^{-2}_v\right)^{-1}\sigma^{-2}_v = \left(1 + \sigma^2_v\sigma^{-2}_{t|t-1}\right)^{-1}. \tag{3.19}$$

But σ_v^2 and $\sigma_{t|t-1}^2$ are both variances and strictly positive, from which we conclude that 0 < K_t < 1. To establish that K_t is bounded away from 1, we first use (3.11) to establish that $\sigma_{t|t}^2 < \sigma_v^2$, so by (3.17), $\sigma_{t|t-1}^2 < \sigma_v^2 + \sigma_w^2$. Thus, $\sigma_v^2\sigma_{t|t-1}^{-2} > \sigma_v^2/\left(\sigma_v^2+\sigma_w^2\right) = 1/\left(1+\sigma_w^2/\sigma_v^2\right) = 1/\left(1+\rho^2\right)$, where $\rho = \sigma_w/\sigma_v$ = the relative migration rate, being the square root of the migration variance as a proportion of the noise variance. With this, we find that $\left(1+\sigma_v^2\sigma_{t|t-1}^{-2}\right) > 1 + 1/\left(1+\rho^2\right) = \left(2+\rho^2\right)/\left(1+\rho^2\right) > 1$. We use this in (3.19) to get $K_t < \left(1+\rho^2\right)/\left(2+\rho^2\right) < 1$. This establishes that K_t is bounded away from 1, as required to establish the exponential decay in j of the weights on y_{t-j} in (3.18), thereby justifying the term "exponentially weighted moving average".

To derive a recursion for K_t, we first substitute (3.17) into (3.11) to obtain the following:

$$\begin{aligned}\sigma_{t|t}^{-2} &= \left(\sigma_{t-1|t-1}^2+\sigma_w^2\right)^{-1} + \sigma_v^{-2} \\ &= \sigma_v^{-2}\left\{\left(\rho^2+\sigma_{t-1|t-1}^2/\sigma_v^2\right)^{-1}+1\right\},\end{aligned}$$

where $\rho = \sigma_w/\sigma_v$, as defined in the previous paragraph. We multiply both sides of this equation by σ_v^2 and recall the definition of K_t, (3.12), to get the following:

$$K_t^{-1} = \left\{\left(\rho^2+K_{t-1}\right)^{-1}+1\right\}. \qquad (3.20)$$

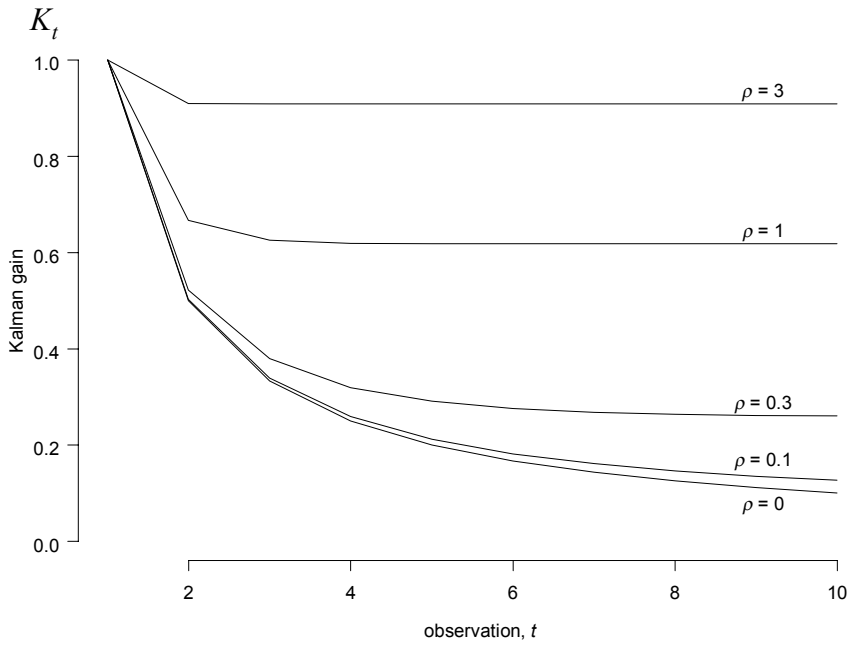
In Figure 3.4, we present the behavior of K_t over time for different levels of ρ assuming $\sigma_{1|0}^2 = \infty$. A similar analysis would establish that $\sigma_{t+1|t}^2$ and $\sigma_{t|t}^2$ also approach asymptotes, which we denote by $\sigma_{\infty+1|\infty}^2$ and $\sigma_{\infty|\infty}^2$ with an obvious abuse of notation.

These latter two asymptotes must satisfy (3.11) and (3.17), which means that $\sigma^2_{\infty+1|\infty} > \sigma^2_{\infty|\infty}$.

This figure suggests that K_t goes to an asymptote, K_∞, say, that depends on ρ. Moreover, for $\rho \geq 0.3$, this asymptote is essentially achieved by observation 10. For $\rho \geq 3$, the asymptote is essentially achieved by $t = 2$. For $\rho \leq 0.1$, the asymptote is not achieved 10 observations. [For $\rho = 0$, we have $K_t = 1/t$, which is easily established using (3.20). Thus, with zero migration variance, the EWMA becomes a straight average, as we would intuitively expect.]

To obtain a formula for this asymptote, we substitute K_∞ for both K_t and K_{t-1} in (3.20) and solve for K_∞. We get the following:

Figure 3.4. Kalman Gain for EWMA vs. Time



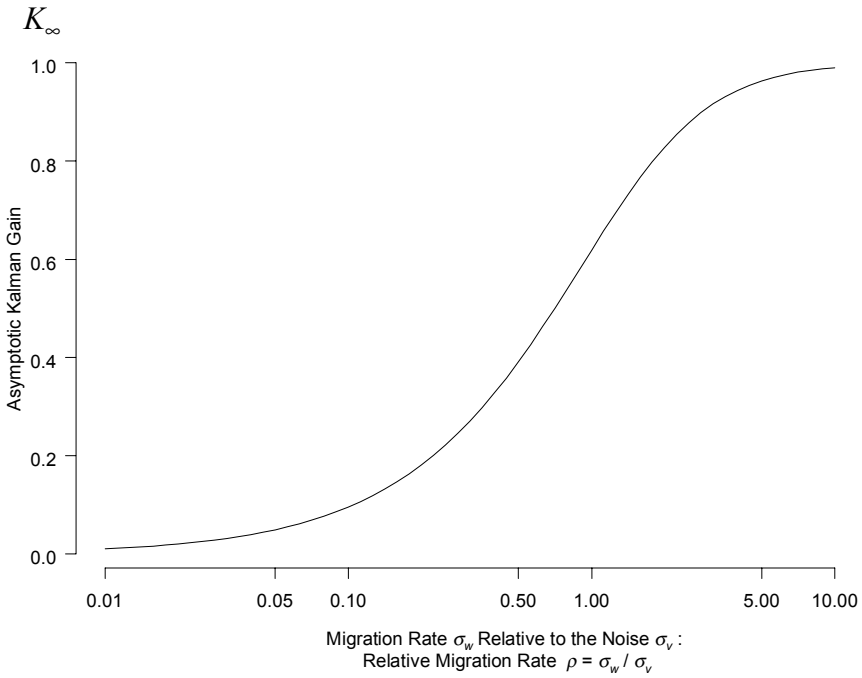
$$\begin{aligned} K_\infty &= \frac{1}{2}\left\{\sqrt{\rho^4 + 4\rho^2} - \rho^2\right\} \\ &= \frac{\rho^2}{2}\left\{\sqrt{1 + \left(4/\rho^2\right)} - 1\right\} \\ &= \rho\left\{\sqrt{1 + \rho^2/4} - (\rho/2)\right\}. \end{aligned} \tag{3.21}$$

To study the asymptotic behavior of K_∞ as ρ gets large or small, we use the binomial theorem as $\sqrt{1+x} = 1 + (x/2) - (x^2/8) + O(x^3)$ in these last two expressions to get the following:

$$\begin{aligned} K_\infty &= 1 - \rho^{-2} + O\left(\rho^{-4}\right) \\ &= \rho\left\{1 - (\rho/2) + \left(\rho^2/8\right) - \left(\rho^4/128\right) + O\left(\rho^6\right)\right\}. \end{aligned} \tag{3.22}$$

The asymptote (3.21) is plotted vs. ρ in Figure 3.5. The most obvious conclusion from (3.21) and Figures 3.4 and 3.5 is that the choice of weight on the last observation is equivalent to specifying ρ = the square root of the migration variance, σ_w^2, as a proportion of the measurement noise, σ_v^2. This relationship quantifies what we would qualitatively expect: With processes that change slowly relatively to the measurement noise, history is more informative that the last observation. On the other hand, rapidly changing processes with relatively informative observations find recent history more relevant than the past for predicting the future. The asymptotic expansions in (3.22) quantify the behavior we see in Figure 3.5 for large and small ρ.

Figure 3.5. Equivalence between EWMA Weight and Relative Migration Rate



Similar expressions can be obtained for $\sigma^2_{\infty+1|\infty}$ and $\sigma^2_{\infty|\infty}$. Moreover, the convergence is monotonic: If $\infty > \sigma^2_{1|0} > \sigma^2_{\infty+1|\infty}$, the convergence is similar to the image in Figure 3.4 but starting some time after $t = 1$. If $\sigma^2_{1|0} < \sigma^2_{\infty+1|\infty}$, the curves *increase* to the asymptote, reflecting the fact that in this case, more information is lost due to the migration (3.1) than is gained from each observation (3.2); see also Kirkendall (1989) and Harvey (1989, pp. 119, 124).

3.5. Bayesian and Traditional Approaches to FIR for EWMA

Lucas and Saccucci (1990) observed that in many applications of a traditional exponentially weighted moving average (EWMA) with a constant weight on the last observations $K_t = K_\infty$ in (3.18), the resulting EWMA may require too many observations

to cross a threshold if the plant is bad, starting with $x_{1|0} = \mu_0$. Their solution is to start with 25, 50 or 75 percent "head start", i.e., with $x_{1|0} = \mu_0 + p(\mu_1 - \mu_0)$, where $p = 0.25$, 0.5, or 0.75.

We believe that Bayesian sequential updating, as outlined in Figures 3.1 and 3.2, provides a more comprehensive and understandable approach to this important problem: In terms of the theory developed in sections 3.2 - 3.4 above, Lucas and Saccucci essentially assume that $\sigma_{1|0} = \sigma_{\infty+1|\infty}$, but that $x_{1|0}$ is misspecified and is better given as $x_{1|0}$ $= \mu_0 + p(\mu_1 - \mu_0)$, where $p = 0.25$, 0.5, or 0.75.

However, a monitor is almost never designed for a situation (plant), that is totally unique. Someone suspects that a fault of a certain type may occur. This suspicion is based on somebody's experience with other applications that bear some resemblance to the problem at hand. This experience provides access to an external reference distribution and data on the reliability of the plant from which estimates for the initial prior ($x_{1|0}$, $\sigma^2_{1|0}$) and migration variance σ^2_w can be derived. We combine this with a gage repeatability and reproducibility study (e.g., NIST 2001, ch. 2) to estimate the noise variance σ^2_v. In this way, Bayesian sequential updating shows a person designing a monitor precisely how to use this relevant information; the previously existing theory is not so clear about the relevance of this information and how to use it.

For the FIR problem, this recommends adjusting $\sigma^2_{1|0}$, not $x_{1|0}$, in the initial prior. The disadvantage is that the weight on the last observation is not constant but must be updated with each observation per (3.20) until convergence to K_∞ is essentially achieved.

Finally, we believe that Bayesian sequential updating, as exemplified by the current work, provides a comprehensive theoretic foundation for work on the short run process control problem, recent discussed, e.g., by Nembhard and Mastrangelo (1998).

3.6. Robustness

Box has noted that robustness is often more important than optimality, since a theoretically optimal solution may be so non-robust that it performs miserably under common discrepancies between reality and standard assumptions. Box and Luceño (1997, pp. 117-127) find that the EWMA provides a quite robust procedure for tracking a drifting process, even if the migration mechanism differs substantially from the random walk of (3.1). This is consistent with other work, e.g, Srivistava and Wu (1993) and Roberts (1966), that finds that the EWMA performs reasonably well under a broad variety of circumstances, though not as well as a cumulative sum in reacting to certain abrupt jumps.

However, a procedure may be robust to one kind of model inadequacy but quite sensitive, nonrobust, to another. For example, an EWMA may follow a process average reasonably well even if the transitions differ substantially from the random walk of (3.1) and the weight on the last observation differs from the optimal. However, we would expect that confidence intervals using the predictive or the prior variance, (3.9) or (3.17), might *not* perform very well if either the measurement or the migration variance, σ_v^2 or σ_w^2, were poorly estimated. Traditional methods for estimating these parameters and

evaluating the applicability of this model are discussed by Box and Luceño (1997, pp. 117-127).

There is at present another practical disadvantage to the use of our Bayesian EWMA, (3.18) with (3.20): We do not currently have a simple method for estimating the run length characteristics for the Bayesian EWMA, other than suggesting that it probably does not differ substantially from the traditional FIR technique proposed by Lucas and Saccucci. However, this is not conceptually a difficult problem and can be addressed, e.g., by Monte Carlo.

3.7. When to Declare a Malfunction?

It is not as easy here to decide when to set an alarm as it is in the discussion of section 2 above, attempting to detect an abrupt jump from good to bad, rather than the gradual drift (random walk) considered here. There, the posterior consisted of one number; here, it is a distribution with two parameters that change over time. One approach might be to develop an appropriate cost structure and develop a decision procedure to minimize the cost per unit time or total discounted cost over an indefinite future. Related problems have discussed by Berger (1985, ch. 7) and West and Harrison (1999, sec. 11.6).

We have not done that here. Instead, we divided the real line into "acceptable", "unacceptable", and "undefined" regions: The plant is "good" if $L \le x_t \le U$, "bad" if $\{x_t \le L_1 < L$ or $x_t \ge U_1 > U\}$, and "undefined" if $\{L_1 < x_t < L$ or $U < x_t < U_1\}$; L and U are "worst acceptable", while L_1 and U_1 are "best unacceptable".

We further simplified the problem by selecting decision limits L^* and U^* and indicating a malfunction when the prior variance $\sigma_{t+1|t}$ is sufficiently small *and* $x_{t+1|t}$ is outside the L^*-U^* limits. The engineering design criteria for this decision procedure (or on-board diagnostic, OBD) are typically expressed in terms of an acceptably small probability of an excessive delay (to detection of x_t being bad) and simultaneously a small probability of a false alarm in the design life of the plant (Box et al. 2000; 2002).

This example is, in essence, an EWMA. Run length distributions of EWMAs have been studied, for example, by Crowder (1987) and Lucas and Saccucci (1990), though the effect of the Bayesian non-constant weights (3.12) and (3.20) seem not to have been described in the literature. Decision limits L^* and U^* could be obtained by Monte Carlo simulation if the work of Crowder and others does not seem appropriate.

3.8. Discussion

In this section, we have derived Bayesian sequential updating for indirect observation of a univariate normal random walk. The result is a Bayesian EWMA, previously discussed by Kirkendall (1989) and Harvey (1989). However, our derivation is, we believe, more methodical and more easily understood and generalized than previous discussions of this case.

Part of this development established that the selection of a weight for the last observation in an EWMA is equivalent to specifying the migration variance σ_w^2 relative to the noise variance σ_v^2, as discussed with (3.20) above. Both of these quantities are generally available from external sources: The migration variance σ_w^2 is tied to

reliability. The noise variance σ_v^2 can be estimated from a metrology study. Moreover, the distribution at the initiation of monitoring $N(x_{1|0}, \sigma_{1|0}^2)$ is obtainable from data typically collected at the end of the production line for manufactured products or from other sources for monitoring in clinical trials or other applications. This places at the disposal of a person designing a monitor relevant information whose use in this context has not been previously discussed in the literature that we have seen. As such, it provides an alternative approach for determining K_t to the integrated moving average estimation procedure recommended by Box and Luceño (1997, sec. 4.8).

As noted in section 3.5, this procedure is essentially as robust as traditional EWMA procedures, being computationally almost identical to them. The Bayesian EWMA provides an additional interpretation as the prior and posterior at each step of objective distributions of units or plants among all with comparable histories. We would not expect this probability interpretation to be robust to departure from serial independence or normality. However, more casual usage of this theory, consistent with current EWMA usage, should be quite robust.

Box and Luceño (1997) comment extensively about an EWMA as a forecast for an integrated moving average IMA(0, 1, 1) process. The present development is asymptotically equivalent to this, and we recommend traditional EWMAs for applications where the transients are unimportant and the situation does not justify the effort of attempting to access other data such as the distribution at first use, gage R & R studies, and reliability data.

The rest of this report generalizes the work of this section. Section 4 develops procedures for effects such as aging that may change the variability as well as the mean.

Sections 5-7 consider multivariate state spaces to support fault isolation. In applications where the output of different sensors should be related, this can allow one sensor to check another, supporting fault isolation without duplicating sensors and increasing per-unit costs. This generalizes the work of Pole, West and Harrison (1994), West (1986), West and Harrison (1986), Gelb (1999), Gordon and Smith (1988, 1990), Harrison and Lai (1999), Lindley and Smith (1972), and others.

REFERENCES

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. (NY: Springer).

Box, G., Graves, S., Bisgaard, S., Van Gilder, J., Marko, K., James, J., Seifer, M., Poublon, M., and Fodale, F. (1999) "Detecting Malfunctions in Dynamic Systems", *Proceedings of the 2000 SAE World Congress & Exposition* (SAE Technical Paper Series number 2000-01-0363).

Box, G., Bisgaard, S., Graves, S., Van Gilder, J., Marko, K., James, F. (2002). "The Waterfall Chart", *Quality Engineering*. To appear.

Box, G., and Luceño, A. (1997) *Statistical Control by Monitoring and Feedback Adjustment* (NY: Wiley).

Crowder, S. V. (1987) "A Simple Method for Studying Run-Length Distributions of Exponentially Weighted Moving Average Charts", *Technometrics*, 29, 401-407.

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

Gelb, A. (1999) *Optimal Applied Estimation* (Cambridge, MA: MIT Press).

Gordon, K., and Smith, A. F. M. (1988) "Modeling and Monitoring Discontinuous Changes in Time Series" in *Bayesian Analysis of Time Series and Dynamic Models*, ed. J. Spall, N.Y.: Marcel Dekker, 359-391.

Gordon, K., and Smith, A. F. M. (1990) "Modeling and Monitoring Biomedical Time Series", *Journal of the American Statistical Association*, 85, 328-337.

Harrison, P. J., and Lai, I. C. H. (1999) "Statistical Process Control and Model Monitoring", *Journal of Applied Statistics*, 26: 273-292.

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

Kalman, R. E. (1960) "A New Approach to Linear Filtering and Prediction Problems", *Journal of Basic Engineering*, 340-345.

Kirkendall, Nancy J. (1989) "The Relationship Between Certain Kalman Filter Models and Exponential Smoothing Models", pp. 89-107 in J. B. Keats and N. F. Hubele, *Statistical Process Control in Automated Manufacturing*, NY: Marcel Dekker.

Lindley, D. V., and Smith, A. F. M. (1972) "Bayes Estimates for the Linear Model", *Journal of the Royal Statistical Society, Series B*, 34, 1-41.

Lucas, J. M., and Saccucci, M. S. (1990) "Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements" (with discussion), *Technometrics*, 32(1), pp. 1-29.

Nembhard, H. B., and Mastrangelo, C. M. (1998) "Integrated Process Control for Startup Operations", *Journal of Quality Technology*, 30(3): 201-211.

NIST (2001) *Engineering Statistics Handbook* (Washington, DC: National Institute of Standards and Technology web-based handbook: http://www.itl.nist.gov/div898/handbook)

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

Roberts, R. W. (1966) "A Comparison of Some Control Chart Procedures", *Technometrics*, 8: 411-430.

Srivistava, M. S., and Wu, Y. (1993) "Comparison of EWMA Cusum, and Shyryayev-Roberts Procedures for Detecting a Shift in the Mean", *Annals of Statistics*, 21: 645-670.

West, M. (1986) "Bayesian Model Monitoring", *Journal of the Royal Statistical Association* B, 48: 70-78.

West, M. and Harrison, P. J. (1986) "Monitoring and Adaptation in Bayesian Forecasting Models", *Journal of the American Statistical Association*, 81, 741-750.

_____ (1999) *Bayesian Forecasting and Dynamic Models*, 2nd ed. (NY: Springer).

4. BAYESIAN EWMA FOR MEAN AND VARIANCE

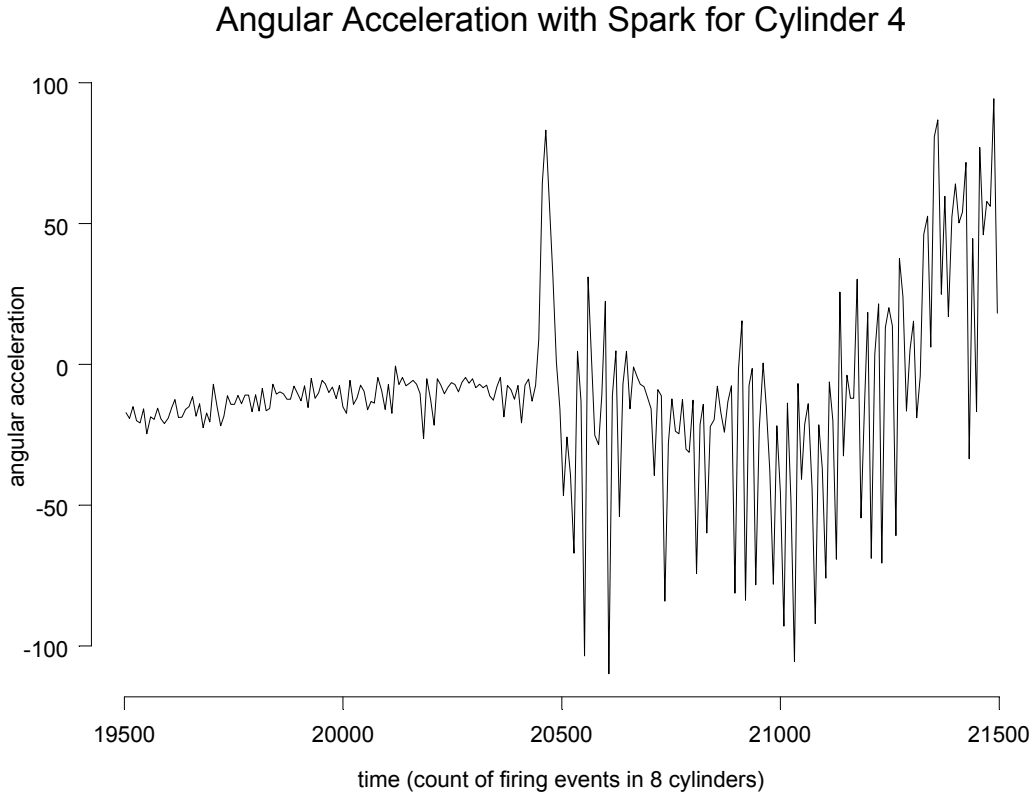
In this section, we consider monitoring a process with drifting mean and variance. We assume the initial prior is normal for the mean and gamma for the precision (i.e., inverse gamma for the variance). The application of Bayesian sequential updating produces exponentially weighted moving averages (EWMAs) for mean and variance. 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. More generally, it provides an alternative model to autoregressive conditional heteroscedasticity (ARCH), popular in the econometrics literature [e.g., Lamoureux and Lastrapes (1990) or Shephard (1994)].

Consider, for example, the plot in Figure 4.1. In this image, a period of relative stability is followed by a time of increased volatility, with an increase in the variability accompanied by wider swings in the apparent central tendency. Data on stock prices sometimes exhibit behavior similar to Figure 4.1: 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 either a gradual or an abrupt increase in the variability. This particular image 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 percent of the way through Figure 4.1, 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 4.1, whereupon the gap between complete and incomplete combustion generates the instabilities we see.

Figure 4.1. A Time Series with Changing Mean and Variability



In this section, we will model these observations with Bayesian exponentially weighted moving averages (EWMAs) for mean and variance. These are fairly simple and elegant tools suitable for many applications where an increase in variability may contribute to the detection of a change. For misfire detection, this model might provide a bridge towards more sophisticated models that could aid in fault isolation as well as

detection. The results of applying these tools to the data of Figure 4.1 appear in Figure 4.2. The EWMA for mean appears in Figure 4.2.1, 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 4.2. Example Smoothing of Mean and Standard Deviation

Figure 4.2.1. Data and Drifting Mean

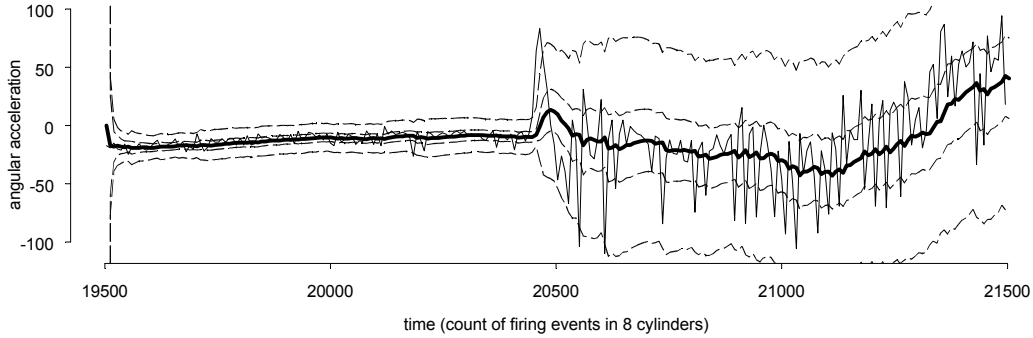
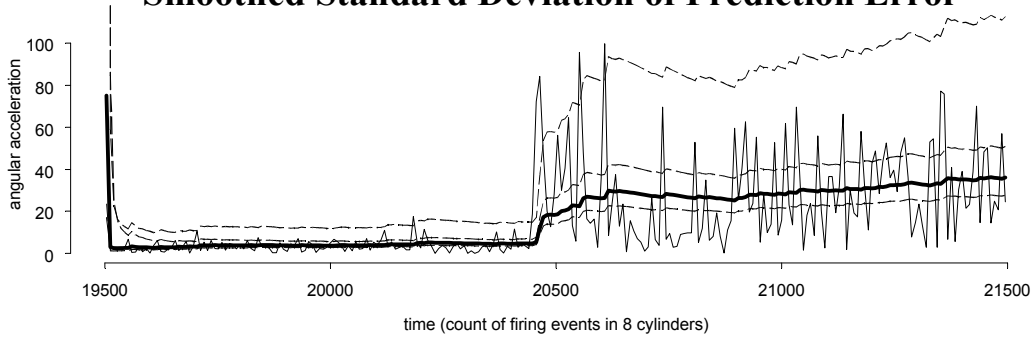


Figure 4.2.2. Absolute Prediction Error and Smoothed Standard Deviation of Prediction Error



Similarly, the dark solid line in Figure 4.2.2 presents the square root of the EWMA for variance of the predictive distribution. It starts large because the substantial 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 percent upper bound for the prediction error.

The theory here follows the development for the normal theory Bayesian EWMA for process mean in section 3 above with the addition of a parameter for relative precision (i.e., reciprocal variance) that evolves following a gamma (or chi-square) distribution conjugate to the normal distributions for process mean. This work is based on more general theory discussed by West and Harrison (1999) and Pole, West and Harrison (1994). These authors present results substantially more general than what we consider here, without clearly relating their "dynamic Bayesian" recursion for precision to an EWMA for variance. We believe also that our double subscript notation, e.g., $x_{t|t-1}$ for a prior and $x_{t|t}$ for a posterior, makes it easier to follow the logic than their superficially simpler notation. [Notation similar to ours was used by Harvey (1989).]

The basic theoretical development is discussed in section 4.1. Confidence intervals needed to evaluate the uncertainty in the various quantities of interest are obtained by integrating out the relative precision, thereby obtaining Student's *t* distributions companion to the normal-gamma pairs. This complicates the theory in exactly the same way that Student's *t* complicates confidence intervals in traditional sampling theory. These complications are discussed separately in section 4.2; they are necessary in applications, but earlier introduction may get in the way of understanding the

primary development of the logic. The application of this theory to the data of Figure 4.1 is discussed in section 4.3. Section 4.4 provides concluding remarks.

The computations described below may seem rather complex to some. Applications requiring only smoothed estimates of mean and variability may need only EWMAs of mean and squared prediction error. 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 4.3 include several steps that merely involve renaming certain quantities. This is done to clarify the theoretical development, making it easier to understand and remember. It has the unfortunate side effect of making the procedure look more complicated than it really is.

4.1. 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 4.1. 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 in section 3 above, 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.

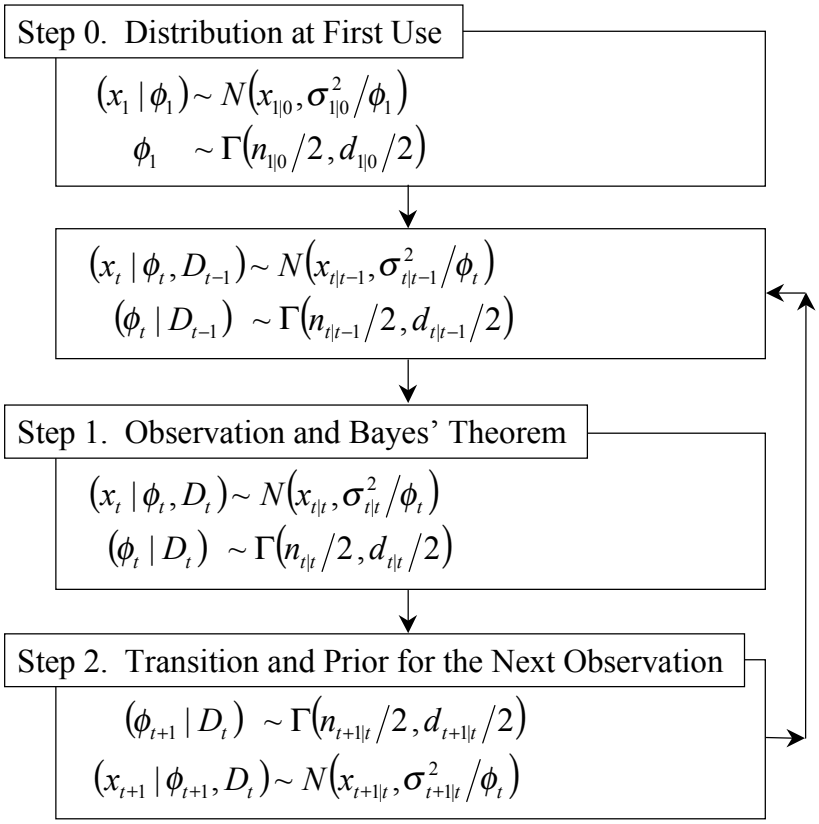
Table 4.1. Algorithm for Bayesian EWMA for Mean and Variance

Assume		
Observation	$y_t = x_t + v_t, \;\; v_t \sim N\left(0, \sigma_v^2/\phi_t\right)$	(4.4)
Migration	$x_{t+1} = x_t + w_t, \;\; w_t \sim N\left(0, \sigma_w^2/\phi_t\right)$	(4.8)
EWMA $x_{t\|t-1}$ for mean x_t		
	$x_{t+1\|t} = \left(1-K_t\right)x_{t\|t-1} + K_t y_t = x_{t\|t-1} + K_t e_t$	(4.17) & (4.23)
Prediction error	$e_t = y_t - x_{t\|t-1}$	(4.16)
Kalman gain	$K_t = 1/\left\{1+1/\left(\rho^2 + K_{t-1}\right)\right\}, \;\; \rho^2 = \sigma_w^2/\sigma_v^2$	(4.14)
EWMA $\tau^2_{t\|t-1}$ for relative variance ϕ_t^{-1}		
	$\tau^2_{t+1\|t} = \left(1-\lambda_t\right)\tau^2_{t\|t-1} + \lambda_t\left(e_t/\sigma_{y\|t-1}\right)^2$	(4.22)
Weight on the last prediction error	$\lambda_t = 1/\left(n_{t\|t-1}+1\right)$	
χ^2 / Student's t degrees of freedom	$n_{t+1\|t} = \delta\left(n_{t\|t-1}+1\right), \;\; 0 < \delta \leq 1$	(4.19) & (4.18)
Prediction error variance	$\sigma^2_{y\|t-1} = \sigma^2_{t\|t-1} + \sigma_v^2$	(4.10)
Prior variance	$\sigma^2_{t+1\|t} = \left(\sigma^{-2}_{t\|t-1} + \sigma_v^{-2}\right)^{-1} + \sigma_w^2$	(4.24) & (4.12)
Confidence interval for x_t via Student's t		
	$\left(x_t \mid D_{t-1}\right) \sim t\left(x_{t\|t-1}, s^2_{t\|t-1}; n_{t\|t-1}\right)$	
Sample variance for the mean	$s^2_{t\|t-1} = \sigma^2_{t\|t-1}\tau^2_{t\|t-1}$	(4.25)
Prediction error via Student's t		
	$\left(e_t \mid D_{t-1}\right) \sim t\left(0, s^2_{y\|t-1}; n_{t\|t-1}\right)$	
Sample variance for prediction error	$s^2_{y\|t-1} = \sigma^2_{y\|t-1}\tau^2_{t\|t-1}$	(4.26)

The theory follows naturally from the two-step Bayesian sequential updating procedure outlined in Figure 4.3.

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. The next step is to modify the resulting posterior to account for a transition in the plant anticipated before the next observation.

Figure 4.3. Bayesian Sequential Updating of Mean and Variance



Step1.0. Model Assumptions. Specifically, at first use, the condition of the plant $(x_1 \mid \phi_1)$ is assumed to follow $N(x_{1|0}, \sigma^2_{1|0}/\phi_1)$, where ϕ_1 = the relative precision, which is assumed to follow a gamma distribution, $\Gamma(n_{1|0}/2, d_{1|0}/2)$. For future reference, we note that $f(x_1, \phi_1) = f(x_1|\phi_1) f(\phi_1)$, where

$$f(x_1|\phi_1) \propto \phi_1^{1/2} \exp\left\{-\frac{\phi_1}{2}\left(\frac{x_1 - x_{1|0}}{\sigma_{1|0}}\right)^2\right\}, \tag{4.1}$$

and

$$f(\phi_1) \propto \phi_1^{(n_{1|0}-2)/2} \exp\{-\phi_1 d_{1|0}/2\}, \tag{4.2}$$

so

$$f(x_1, \phi_1) \propto \phi_1^{(n_{1|0}-1)/2} \exp\left\{-\frac{\phi_1}{2}\left[\left(\frac{x_1 - x_{1|0}}{\sigma_{1|0}}\right)^2 + d_{1|0}\right]\right\}. \tag{4.3}$$

(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

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

so

$$f(y_t|x_t, \phi_t) \propto \phi_t^{1/2} \exp\left\{-\frac{\phi_t}{2}\left(\frac{y_t - x_t}{\sigma_v}\right)^2\right\}. \tag{4.4}$$

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

$$f(x_t, \phi_t|D_{t-1}) = f(x_t|\phi_t, D_{t-1}) f(\phi_t|D_{t-1}),$$

where

$$f(x_t|\phi_t, D_{t-1}) \propto \phi_t^{1/2} \exp\left\{-\frac{\phi_t}{2}\left(\frac{x_t - x_{t|t-1}}{\sigma_{t|t-1}}\right)^2\right\}, \tag{4.5}$$

and

$$f(\phi_t|D_{t-1}) \propto \phi_t^{(n_{t|t-1}-2)/2} \exp\{-\phi_t d_{t|t-1}/2\}, \tag{4.6}$$

so

$$f(x_t, \phi_t|D_{t-1}) \propto \phi_t^{(n_{t|t-1}-1)/2} \exp\left\{-\frac{\phi_t}{2}\left[\left(\frac{x_t - x_{t|t-1}}{\sigma_{t|t-1}}\right)^2 + d_{t|t-1}\right]\right\}. \tag{4.7}$$

When $t = 1$, (4.5) - (4.7) is provided by the initial prior (4.1) - (4.3); later, it is provided by the output of step 2, transition, after the previous observation.

As outlined in Figure 4.3, 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 $\{N(x_{t|t},\ \sigma^2_{t|t}/\phi_t),\ \Gamma(n_{t|t}/2,\ d_{t|t}/2)\}$ and $\{N(x_{t+1|t},\ \sigma^2_{t+1|t}/\phi_{t+1}),\ \Gamma(n_{t+1|t}/2,\ d_{t+1|t}/2)\}$, respectively.

In particular, the transition in location is modeled as

$$x_t = x_{t-1} + w_{t-1}, \quad w_t \sim N(0,\ \sigma^2_w/\phi_t), \tag{4.8}$$

while the distribution for the relative precision is discounted from $\Gamma(n_{t|t}/2,\ d_{t|t}/2)$ for ϕ_t to $\Gamma(n_{t+1|t}/2,\ d_{t+1|t}/2)$ for ϕ_{t+1}, with $n_{t+1|t} = \delta\, n_{t|t}$ and $d_{t+1|t} = \delta\, d_{t|t}$, for some δ with $0 < \delta \leq 1$.

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 Variance, and (1.1c) Kalman Gain, as we now explain.

Step 1.1a. Predictive Distribution. We want the marginal distribution $(y_t \mid \phi_t,\ D_{t-1})$. From (4.4), we see that y_t is the sum of two independent, normally distributed random variables, so $(y_t \mid \phi_t,\ D_{t-1})$ is also a normal distribution, with mean and variance being the sums of the means and variances of x_t and v_t:

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

where

$$f_t = x_{t|t-1}, \tag{4.9}$$

and

$$\sigma^2_{y|t-1} = \sigma^2_{t|t-1} + \sigma^2_v . \tag{4.10}$$

[Note that ϕ_t^{-1} is a common factor of the variances of $(x_t \mid \phi_t, D_{t-1})$ and the observation per (4.4), and is therefore also a common factor of the predictive variance.]

Step 1.1b. Posterior Variance. When a quantity with a normal prior is observed with additive normal error, the posterior is also normal. The posterior mean is a weighted average of the prior mean and the observation with weights inversely proportional to the variances, and the posterior variance is the reciprocal sum of the reciprocal variances (e.g., DeGroot 1970, p. 167). We shall write this as follows:

$$(x_t \mid \phi_t, D_t) \sim N(x_{t|t},\ \sigma^2_{t|t}/\phi_t),$$

where

$$x_{t|t} = \frac{\sigma^{-2}_{t|t-1} x_{t|t-1} + \sigma^{-2}_v y_t}{\sigma^{-2}_{t|t-1} + \sigma^{-2}_v} = \sigma^2_{t|t} \left\{ \sigma^{-2}_{t|t-1} x_{t|t-1} + \sigma^{-2}_v y_t \right\}, \tag{4.11}$$

and

$$\sigma^{-2}_{t|t} = \sigma^{-2}_{t|t-1} + \sigma^{-2}_v . \tag{4.12}$$

A squared reciprocal scale factor is sometimes called a "precision" (e.g., DeGroot 1970, p. 38); for $N(\mu,\ \sigma^2)$, the precision is σ^{-2}. With this concept, (4.12) says that the posterior precision, $\phi_t \sigma^{-2}_{t|t}$, is the sum of the precisions of the prior $\phi_t \sigma^{-2}_{t|t-1}$ and the observation $\phi_t \sigma^{-2}_v$; the relative precision ϕ_t cancels leaving (4.12).

Step 1.1c. Kalman Gain. The weight on the last observation y_t in (4.11) is called the Kalman gain, and will be denoted as follows:

$$K_t = \sigma^2_{t|t} \sigma^{-2}_v . \tag{4.13}$$

A careful analysis of K_t reveals that it is dimensionless, depending essentially on the ratio of the migration variance to the observation variance. Denote this ratio by $\rho^2 = \sigma_w^2/\sigma_v^2$. Since the relative precision ϕ_t cancels, the analysis of section 3.4 above applies, giving us the following simple recursion for K_t:

$$K_t^{-1} = \left\{\left(\rho^2 + K_{t-1}\right)^{-1} + 1\right\}. \tag{4.14}$$

To use K_t in (4.11), we first use (4.12) to obtain

$$\sigma_{t|t-1}^{-2} = \sigma_{t|t}^{-2} - \sigma_v^{-2}.$$

We substitute this with (4.13) into (4.11) to get

$$\begin{aligned} x_{t|t} &= \sigma_{t|t}^2 \left\{\left(\sigma_{t|t}^{-2} - \sigma_v^{-2}\right)x_{t|t-1} + \sigma_v^{-2} y_t\right\} \\ &= x_{t|t-1} + K_t\left(y_t - x_{t|t-1}\right). \end{aligned} \tag{4.15}$$

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. 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 compute the (*a*) prediction error and use that to update (*b*) the posterior mean and (*c*) the posterior precision.

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

$$e_t = y_t - f_t. \tag{4.16}$$

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 (4.15), as

$$x_{t|t} = x_{t|t-1} + K_t e_t . \tag{4.17}$$

Step 1.2c. Posterior Precision. We now combine the predictive distribution (4.9) - (4.10) with the prior for the common precision (4.6) as

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

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

where

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

and (4.18)

$$d_{t|t} = (e_t / \sigma_{y|t-1})^2 + d_{t|t-1} .$$

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

This completes step 1, observation, in Bayesian sequential updating as outlined in Figure 4.3. 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 $(x_t \mid \phi_t, D_t)$ and $(\phi_t \mid D_t)$ from step 1, we consider the transition for the relative precision ϕ_t to ϕ_{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 δ, with $0 < \delta \leq 1$, so

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

where

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

and

$$d_{t+1|t} = \delta\, d_{t|t}. \tag{4.19}$$

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

$$\phi_{t+1} = \gamma_t \phi_t / \delta\,, \tag{4.20}$$

where $(\gamma_t|D_t) \sim \text{Beta}[\delta n_{t|t}/2,\, (1 - \delta) n_{t|t}/2]$; for this distribution, $E(\gamma_t|D_t) = \delta$, so $E(\phi_{t+1}|D_t) = \phi_t$.

It may help to understand δ to note that [using (4.18) and (4.19)]

$$n_{t|t} \to 1/(1-\delta) \text{ as } t \to \infty \tag{4.21}$$

(West and Harrison, p. 362). Thus, selecting δ is equivalent to specifying the degrees of freedom in the steady-state chi-square distribution for the relative precision.

Expression (4.20) 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 4.1 of a change in throttle angle. However, with or without (4.20) and possible refinements, we must still ignore the difference between ϕ_t and ϕ_{t+1} in modeling the transition from x_t to x_{t+1}, which is the subject of step 2.2.

Expressions (4.18) - (4.19) are equivalent to an EWMA for the relative variance, which we define as $\tau^2_{t+1|t} = d_{t+1|t}/n_{t+1|t}$. By (4.19) and (4.18), this is

$$\tau^2_{t+1|t} = \frac{\delta\left(d_{t|t-1} + e_t^2/\sigma^2_{y|t-1}\right)}{\delta\left(n_{t|t-1}+1\right)} = \frac{n_{t|t-1}\tau^2_{t|t-1} + e_t^2/\sigma^2_{y|t-1}}{n_{t|t-1}+1} = \tau^2_{t|t},$$

so

$$\tau^2_{t+1|t} = (1-\lambda_t)\tau^2_{t|t-1} + \lambda_t\, e_t^2/\sigma^2_{y|t-1},$$

where (4.22)

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

We will use this to evaluate variability in section 4.2 not conditioned on the unknown relative precision ϕ_t.

Step 2.2. Prior Mean and Variance. Given the posterior $(x_t \mid \phi_t, D_t)$ from step 1, with the transition (4.8), we get $(x_{t+1} \mid \phi_t, D_t) \sim N(x_{t+1|t},\ \sigma^2_{t+1|t}/\phi_t)$, where

$$x_{t+1|t} = x_{t|t}, \tag{4.23}$$

and

$$\sigma^2_{t+1|t} = \sigma^2_{t|t} + \sigma^2_w. \tag{4.24}$$

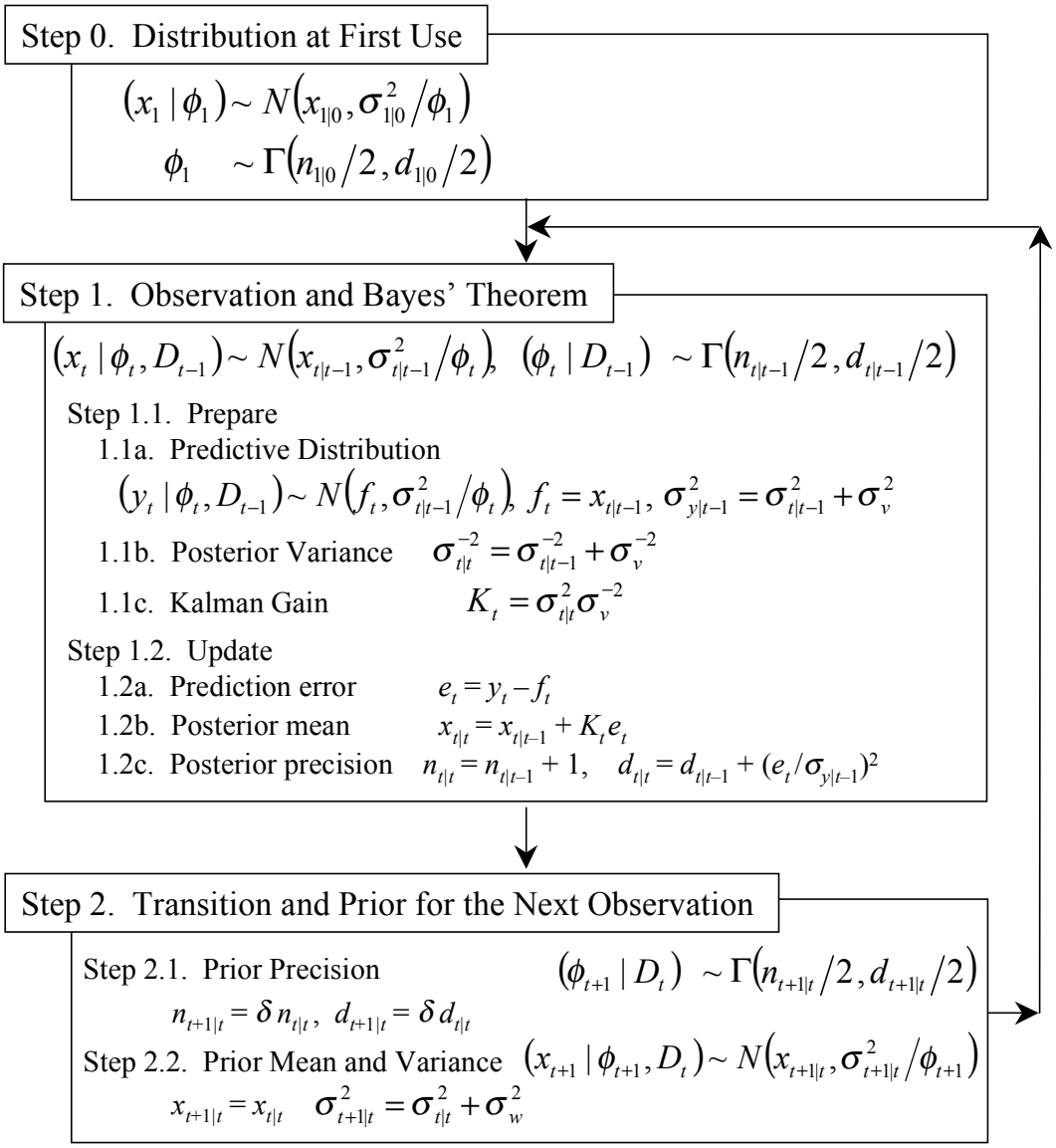
When we return to step 1 for the next observation, we replace ϕ_t with ϕ_{t+1}. If $\delta = 1$, the relative precision is assumed to be constant, so $\phi_{t+1} = \phi_t$, and this step is obvious. If $\delta < 1$, it is not clear (at least to the present authors) that our model assumptions necessarily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

This completes step 2. The resulting prior output from one point in time $\{N(x_{t+1|t}, \sigma^2_{t+1|t}/\phi_{t+1}),\ \Gamma(n_{t+1|t}/2,\ d_{t+1|t}/2)\}$ becomes an input prior for step 1.1, $\{N(x_{t|t-1},\ \sigma^2_{t|t-1}/\phi_t)$,

$\Gamma(n_{t|t-1}/2,\ d_{t|t-1}/2)\}$, at the next point in time. In this way, observations are processed sequentially as they arrive. If the model (4.1) - (4.8) is correct, then the prior $\{N(x_{t+1|t},\ \sigma^2_{t+1|t}/\phi_t),\ \Gamma(n_{t+1|t}/2,\ d_{t+1|t}/2)\}$ summarizes all the information in $D_t = \{y_t, y_{t-1}, ..., y_1, x_{1|0}, \sigma_{1|0}\}$ about the state of the plant at time $t+1$ [ignoring the approximation involved in replacing ϕ_t by ϕ_{t+1} following (4.23) - (4.24)].

The most important expressions in this section are summarized in Figure 4.4. Unfortunately, these results can not be used directly, because they are all conditioned upon the unknown relative precision ϕ_t. We next integrate out this unknown precision, obtaining Student's *t* marginals (section 4.2). The results are applied to the data of Figure 4.1 in section 4.3. We then consider the asymptotic behavior of the smoothing parameters in section 4.4.

Figure 4.4. Bayesian EWMA Normal-Gamma Iteration



4.2. Student's *t* Confidence Intervals

The results of section 4.1 can not be used directly, because they involve the unknown relative precision ϕ_t. To apply the results, we must integrate out ϕ_t, obtaining Student's *t* marginals from normal-gamma distributions discussed in section 4.1.

The normal-gamma density as in (4.3) and (4.7) is as follows:

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

We integrate out ϕ to get the following:

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

where

$$s^2 = \sigma^2\tau^2 \text{ and } \tau^2 = d/n.$$

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

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

Applying this to the prior (4.7) gives us

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

with (4.25)

$$s_{t|t-1} = \sigma_{t|t-1}\,\tau_{t|t-1},$$

and $\tau^2_{t|t-1}$ is the relative variance computed recursively by the EWMA in (4.22). This was used in Figure 4.2.1 to determine the inner set of dashed lines as a 99.7 percent confidence interval for $(x_{t+1} \mid D_t)$.

Similarly, for the predictive distribution not conditioned on the unknown relative precision ϕ_t, we get the following from (4.9)-(4.10) and (4.6):

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

with (4.26)

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

This was used to compute the outer dashed lines in Figures 4.2.1 and 4.2.2. The inner pair of dashed lines in Figure 4.2.2 utilize the (0.0015 and 0.9985) quantiles of the gamma distribution of (4.19) to place confidence limits on $s_{y|t-1}$.

We now discuss the computation of a few of the numbers plotted in Figure 4.2.

4.3. Sample Computations

The sample computations described in this section may be 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 difference is visible in the different confidence intervals associated with the three concepts. In Figure 4.2, 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. 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 4.1.

In the present model, $\text{var}(y_t \mid \phi_t) = \sigma_v^2/\phi_t$ per (4.4) and $\text{var}(x_t \mid \phi_t) = \sigma_w^2/\phi_t$ per (4.8). Since there are no other constraints on ϕ_t, we can without loss of generality set σ_v^2

= 1. With this choice, ϕ_t becomes the observation precision, and σ_w^2 = $\text{var}(x_t \mid \phi_t)/\text{var}(y_t \mid \phi_t) = \rho^2$ = the migration variance as a proportion of the observation variance.

With this choice, table 4.2 begins by recording that the observation parameters for variance σ_v^2 and precision σ_v^{-2} are both 1. Similarly, table 4.2 reports that the migration variance parameter σ_w^2 is assumed to be 0.01, while the variance discount factor $\delta = 0.98$. As discussed in section 3 above, the migration variance σ_w^2 is related to reliability and is equivalent to specifying the degree of smoothing, while δ = 0.98 corresponds to an asymptotic degrees of freedom of 50 per (4.21), one fifth of the observations in Figures 4.1 and 4.2.

The first row in the body of the table gives the first three observations y_t plotted in Figure 4.1. The initial prior for the mean per (4.1) is specified in terms of the mean $x_{1|0}$ and standard deviation $\sigma_{1|0}$. A rough estimate obtained simply by looking at Figure 4.1 is $x_{1|0} = 0$ and $\sigma_{1|0} = 25$; this latter number means that $\sigma_{1|0}^{-2} = 0.0016$ and $\sigma_{1|0}^2 = 625$. For later observations, $x_{t|t-1}$, $\sigma_{t|t-1}^{-2}$, and $\sigma_{t|t-1}^2$ are taken from step 2.2 at the bottom of Table 4.2. We carry both the precision and the variance in Table 4.2 because Bayes' theorem for the normal distribution tells us to add precisions, while a simple sum of random variable requires addition of variances.

Table 4.2. Illustrative Calculations for Bayesian Sequential Updating

Observation variability:

(4.4) conditional variance $\sigma_v^2 = 1$; conditional precision $\sigma_v^{-2} = 1$

Transition:

(4.8) conditional migration variance $\sigma_w^2 = 0.01$; (4.19) variance discount factor $\delta = 0.98$

Step					
	time		1	2	3
1. Observation and Bayes' Theorem					
Observation: Measured angular acceleration		$y_t =$	-17.108	-19.095	-14.985
1.0. Prior (4.1) - (4.3), (4.5) - (4.7)					
mean		$x_{t\|t-1} =$	0.000	-17.081	-18.092
conditional precision		$\sigma_{t\|t-1}^{-2} =$	0.0016	0.992	1.953
variance		$\sigma_{t\|t-1}^{2} =$	625.000	1.008	0.512
EWMA for variance		$\tau_{t\|t-1}^{2} =$	9.000	4.734	3.817
degrees of freedom		$n_{t\|t-1} =$	1.000	1.960	2.901
sample standard deviation (4.25)		$s_{t\|t-1} =$	75.000	2.185	1.398
Student's t $\alpha = 0.0015$			212.205	19.080	9.316
limits for	upper		15915.35	24.606	-5.067
the mean	lower		-15915.35	-58.767	-31.117
1.1. Prepare					
1.1a. Predictive distribution					
mean (4.9)		$f_t =$	0.000	-17.081	-18.092
conditional variance (4.10)		$\sigma_{y\|t-1}^{2} =$	626.000	2.008	1.512
sample standard deviation (4.26)		$s_{y\|t-1} =$	75.060	3.083	2.402
99.7% confidence interval					
Student's t $\alpha = 0.0015$			212.205	19.080	9.316
for the	upper		15928.10	41.750	4.290
observation	lower		-15928.10	-75.912	-40.474
for (absolute prediction error $\|e_t\|$) =			15928.10	58.831	22.382
for sample standard deviation $s_{y\|t-1}$					
$s_{0.0015}^2 = \chi^2(0.0015; n_{t\|t-1}) / n_{t\|t-1} =$			3.53×10^{-6}	1.33×10^{-3}	1.54×10^{-3}
$s_{0.9985}^2 = \chi^2(0.9985; n_{t\|t-1}) / n_{t\|t-1} =$			10.079	6.582	6.487
upper $s_{y\|t-1} / s_{0.0015}$			39926.11	84.550	61.279
lower $s_{y\|t-1} / s_{0.9985}$			23.643	1.202	0.943

1.1b. Posterior variability (4.12)				
conditional precision	$\sigma_{t\|t}^{-2}$ =	1.0016	1.992	2.953
conditional variance	$\sigma_{t\|t}^{2}$ =	0.998	0.502	0.339
1.1c. Kalman gain (4.13)	K_t =	0.998	0.502	0.339
1.2. Update				
1.2a. Prediction error (4.16)	e_t =	-17.108	-2.014	3.107
standardize squared prediction error	$\left(e_t^2/\sigma_{y\|t-1}^2\right)$ =	0.468	2.020	6.384
log(likelihood)		-5.514	-2.460	-2.768
1.2b. Posterior mean (4.17)	$x_{t\|t}$ =	-17.081	-18.092	-17.040
1.2c. Posterior relative precision				
degrees of freedom (4.18)	$n_{t\|t}$ =	2.000	2.960	3.901
weight on squared prediction error (4.22)	λ_t =	0.500	0.338	0.256
deviation of standardized squared prediction error from prior relative variance	$\left[\left(e_c^2/\sigma_{y\|t-1}^2\right)-\tau_{t\|t-1}^2\right]$ =	-8.532	-2.713	2.567
EWMA for variance (4.22)	$\tau_{t\|t}^2$ =	4.734	3.817	4.475

2. Transition and prior for the next observation

2.1. Prior precision				
EWMA for variance (4.22)	$\tau_{t+1\|t}^2$ =	4.734	3.817	4.475
degrees of freedom (4.19)	$n_{t+1\|t}$ =	1.960	2.901	3.823
2.2. For process average				
mean (4.23)	$x_{t+1\|t}$ =	-17.081	-18.092	-17.040
conditional variance (4.24)	$\sigma_{t+1\|t}^2$ =	1.008	0.512	0.349
precision	$\sigma_{t+1\|t}^{-2}$ =	0.992	1.953	2.868
sample standard deviation	$s_{t+1\|t}$ =	2.185	1.398	1.249

Similarly, the initial value of the EWMA for variance $\tau_{1|0}^2$ was chosen as 3^2 just by eying Figure 4.1. 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. Thus, $\tau_{1|0}^2 = 9$. As for the prior for the mean, for $t > 1$, we get $\tau_{t|t-1}^2$ and $n_{t|t-1}$, from step 2.1 at the bottom of Table 4.2. Next, the sample standard deviation for x_t, $s_{t|t-1}$, is computed per (4.25) as the square root of the product of the EWMA for variance and the conditional variance parameter per (4.25), 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 4.2. This produces the inner pair of dashed lines in Figure 4.2.1.

With the prior specified for each step, we now proceed as outlined in Figure 4.4 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 predictive conditional variance parameter $\sigma^2_{y|t-1}$ is the sum of the prior and observation variance parameters, which produces 626 for the first observation. The square root of the product of the EWMA for variance $\tau^2_{t|t-1}$ and the predictive conditional variance parameter $\sigma^2_{y|t-1}$ 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 4.2.2.

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 4.2.1 and 4.2.2. A confidence interval for the predictive sample standard deviation is obtained by referring it to a chi-square distribution. This produces the inner pair of dashed lines in Figure 4.2.2.

We now proceed to step 1.1b, computing the posterior conditional precision parameter as the sum of the prior and observation precision parameters. The conditional

posterior variance parameter is the reciprocal of the corresponding precision parameter. Next, in step 1.1c we compute the weight on the last observation, the Kalman gain, which per (4.13) is the posterior variance times the observation precision. 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 prior precision 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 $\sigma_w^2 = 0.01$ 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 $\rho^2 = \sigma_w^2/\sigma_v^2 = 0.01$, the Kalman gain continues down to an asymptote at 0.0951; see (3.21) in section 3.4 above. 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 (4.16), step 1.2a. This is used to update both the EWMA for mean and for variance. To prepare for updating the EWMA for variance, we square this and divide by its corresponding conditional 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, and it is important to note that the appropriate likelihood in this case rests on the marginal

predictive distribution for the next observation, after integrating out the unknown relative precision ϕ_t.

Step 1.2b, (4.17), 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 (4.22). 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 $\tau^2_{t|t} = 4.734$ for $t = 1$. This completes step 1.

It remains to modify the posterior to account for anticipated migration prior to the next observation. Per (4.22), $\tau^2_{t+1|t} = \tau^2_{t|t}$. However, the degrees of freedom are discounted by the factor δ. For $t = 1$ with $\delta = 0.98$ this discounts the posterior degrees of freedom from 2 to 1.96 for the prior at $t = 2$. Per (4.23), 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 prior variance parameter is the posterior plus migration variance parameters, giving us 1.008 for $t = 1$. We reciprocate this to get the prior precision parameter. 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 conditional variance parameter, which is $s_{t+1|t} = 2.185$ for $t = 1$.

4.3. Discussion

We believe that this discussion of a Bayesian EWMA for mean and variance provides another illustration of the power of Bayesian sequential updating for designing monitors. Other procedures for monitoring mean and variance have previously appeared in the literature, but without such obvious ties to a unifying principle for designing monitors. 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.

REFERENCES

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.

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).