

# Robustification of Kalman Filter Models

RICHARD J. MEINHOLD and NOZER D. SINGPURWALLA*

Kalman filter models based on the assumption of multivariate Gaussian distributions are known to be nonrobust. This means that when a large discrepancy arises between the prior distribution and the observed data, the posterior distribution becomes an unrealistic compromise between the two. In this article we discuss a rationale for how to robustify the Kalman filter. Specifically, we develop a model wherein the posterior distribution will revert to the prior when extreme outlying observations are encountered, and we point out that this can be achieved by assuming a multivariate distribution with Student-$t$ marginals. To achieve fully robust results of the kind desired, it becomes necessary to forsake an exact distribution-theory approach and adopt an approximation method involving "poly-$t$" distributions. A recursive mechanism for implementing the multivariate-$t$–based Kalman filter is described, its properties are discussed, and the procedure is illustrated by an example.

KEY WORDS: Automatic control; Bayes law; Bounded influence functions; Kalman filtering; Multivariate Student-$t$ distributions; Non-Gaussian filtering; Poly-$t$ densities; Robustness; Signal processing.

## 1. INTRODUCTION

The Kalman filter (KF) model, successfully used by engineers, economists, and other scientists, has come to be regarded with increasing interest by statisticians; for example, see Harrison and Stevens (1976), West, Harrison, and Migon (1985), Diderrich (1985), Meinhold and Singpurwalla (1983, 1987), and the references therein.

The model—which stems from Wiener's (1949) theory for prediction and smoothing—relates a sequence of observations $Y_1, Y_2, \ldots, Y_t$ to a set of unobservable quantities $\theta_1, \theta_2, \ldots, \theta_t$ via the *observation equation*

$$Y_t = F_t\theta_t + v_t, \tag{1.1}$$

where the evolution of $\theta_t$ is described by the *system equation*

$$\theta_t = G_t\theta_{t-1} + w_t, \tag{1.2}$$

with $F_t$ and $G_t$ (which may be scalars or matrices, depending on the dimensions of $Y_t$ and $\theta_t$) assumed known. The innovations $\{v_t\}$ and $\{w_t\}$ have mean 0 and are assumed serially and pairwise independent with known covariance matrices $V_t$ and $W_t$, respectively. Under the special case of a Gauss–Markov model, posterior to time $t-1$, but prior to $t$, $\theta_{t-1}$ is assumed Gaussian with mean $\hat{\theta}_{t-1}$ and variance $\Sigma_{t-1}$. Upon the receipt of $Y_t$, and assuming that $v_t$ and $w_t$ are also Gaussian, Bayes's law is used to show that $\theta_t$ is Gaussian with mean and variance

$$\hat{\theta}_t = G_t\hat{\theta}_{t-1} + K_t(Y_t - F_tG_t\hat{\theta}_{t-1}) \tag{1.3}$$

and

$$\Sigma_t = (I - K_tF_t)R_t, \tag{1.4}$$

respectively, where $R_t = G_t\Sigma_{t-1}G_t' + W_t$ and $K_t = R_tF_t'(F_tR_tF_t' + V_t)^{-1}$; $F_t'$ denotes the transpose of $F_t$ and $F_tG_t\hat{\theta}_{t-1}$ is the predicted value of $Y_t$, prior to time $t$.

Simplicity and an adaptive nature are attractive features of this recursive scheme [which may also be derived via a least squares approach, as in Kalman (1960)]. On the other hand, the model is *nonrobust:* The mean (1.3) is an unbounded function of the discrepancy between $Y_t$ and its prior predicted value $F_tG_t\hat{\theta}_{t-1}$, whereas the variance (1.4) does not depend on the observed data. Thus a spurious outlying value of $Y_t$ would adversely affect inference about $\theta_t$. The aim of this article is to "robustify" the KF—that is, to consider the treatment of such outliers in a manner that will produce a robust model, entirely within the strictures of the Bayesian paradigm. By robustification, we mean a "judicious and grudging *elaboration of the model* to ensure against particular hazards" (Box 1980, p. 396). Furthermore, we wish to achieve the desired protection through a fully Bayesian approach.

Previous approaches to robustifying the KF were undertaken by Masreliez (1975), Masreliez and Martin (1975, 1977), Morris (1976), Ershov and Lipster (1978), West (1981), Tsai and Kurz (1983), Boncelet and Dickinson (1984), and Kitagawa (1987). These efforts, however, were guided by sampling-theory principles (e.g., Hampel, Ronchetti, Rousseeuw, and Stahel 1986); in what follows, we refer to this point of view as *conventional robustness*. The attitude of Sorensen and Alspach (1971), Martin (1979), West (1981, 1982), Smith and West (1983), and Guttman and Pena (1985) is more in tune with ours; they proceeded along distribution-theoretic lines by considering non-Gaussian densities.

## 2. BAYESIAN PERSPECTIVES ON ROBUSTNESS AND THEIR RELEVANCE TO KF MODELS

The notion of robustness in a Bayesian setting has been considered by many authors (e.g., see Berger 1984). In general, what is meant by robustness—and a specification of criteria for identifying when lack of robustness can be a problem—can vary considerably for different settings. In the setup of this article, we assume that the structural part of the KF model provides an appropriate view of reality; accordingly, an attempt at robustification will not entertain changes in the basic structure given in (1.1) and

* Richard J. Meinhold is Research Associate, Institute for Reliability and Risk Analysis, and Nozer D. Singpurwalla is Professor, Departments of Operations Research and Statistics, George Washington University, Washington, DC 20052. This work was supported by Office of Naval Research Contract N00014-85-K-0202 (Project NR 347-128-410) and U.S. Army Research Office Grant DAAL 03-87-K-0056. The authors gratefully acknowledge the excellent comments of the associate editor and two referees.

(1.2). Also, we follow a guideline that stems from De Finetti's (1961) view that "a proper Bayesian effort will not be concerned with the behavior of particular estimates but rather with that of the entire posterior distribution" (p. 203). Therefore, we need to decide what should happen to the posterior distribution of $\theta_t$ when a discrepancy between the prior specification and the observed data arises. This question has been considered in a more general context, and two schools of thought have emerged:

1. One school, advocated by Lindley (1968) and Leonard (1974), argues that the data should be emphasized in preference to the prior (also see O'Hagan 1987). This idea is in line with the notion of the prior "washing out" as more data are accumulated.

2. The second school, advocated by De Finetti (1961), *parallels* the conventional robustness argument that the specified prior distribution embodies an educated expectation of what should happen, so an observation that deviates markedly from its prediction should be regarded as suspicious and given less weight in the formation of the posterior.

In the KF situation, the prior at time $t$ will have evolved from $t - 1$ preceding observations, and the data will consist of only one observed value. Therefore, the more meaningful point of view (and the one that we adopt), would be that of De Finetti: A robust KF model would be one for which the posterior distribution of the state of nature would return to its prior as the observation departs significantly from its predicted value.

Dawid (1973) (also Hill 1974; Meeden and Isaacson 1977) formalized and unified the two schools of thought by establishing that the behavior of either type can be attained through the appropriate specification of the probability models. That is, when the prior is weak (strong), as measured by its having a heavier (lighter) tail than the likelihood, the posterior will converge to the likelihood (prior) when the prior and the likelihood diverge. Under the familiar Gaussian assumptions of the nonrobust KF, both the prior and the likelihood have identical tails, so neither dominates the other and the posterior turns out to be a compromise between the prior and the likelihood; this may be seen by rewriting the posterior mean (1.3) as $\hat{\theta}_t = (I - K_tF_t)G_t\hat{\theta}_{t-1} + K_tY_t$. This suggests that an attempt at robustification may be pursued by leaving the basic KF model unaltered, but replacing random Gaussian variables with those having distributions that provide for a wider variety of behavior of tail-area probabilities. This is in line with earlier work on robustness in statistics, in which notions of "outlier prone" and "outlier resistant" distributions, as considered by Green (1974, 1976) and Neyman and Scott (1971), became relevant (also see Gather and Kale 1986). The need to replace the Gaussian distribution with one that is symmetric and behaves much like the Gaussian in the central area (but has heavier tails) leads to a consideration of the Student-$t$ distribution (e.g., West 1982; Zellner 1976). Accordingly, we examine the merits of specifying that $\{v_t\}$ and $\{w_t\}$ have marginal Student-$t$ distributions, with the joint distributions across time either being or not being a multivariate Student-$t$ distribution. The former yields a measure of location that is nonrobust, but it has the property that the scale parameter depends on the observed data (see Sec. A.1 in the Appendix). The latter, which leads to a desirable form of robustification, is discussed in Section 3.

Before proceeding further, it may be worthwhile to remark that in assuming that the innovation series follows a Student-$t$ distribution, there is no implication that the analyst indeed believes that this is truly the case, any more than there is reason to believe that any phenomenon is actually Gaussian (see Anderson and Moore 1979, p. 10). Rather, the distributional assumptions are regarded as reasonable approximations to reality, tempered by concern for the adverse consequences caused by spurious (extremely large or small) observations. One may wish to contrast this with the robustness point of view taken by Berger (1984), concerns about which can be found in the accompanying discussions by Hill, Kadane, and Lindley.

## 3. THE KF MODEL WITH INDEPENDENT STUDENT-$t$ DISTRIBUTED VARIABLES

As stated before, Dawid's theorem provides a general framework for robustification in a distribution-theoretic framework. His result holds for unspecified distributions, and it pertains to necessary conditions for convergence in distribution of the posterior to the prior; these conditions are more restrictive than necessary for the KF model. The multivariate-$t$ distribution (see Theorem A.1 in the Appendix) violates one of Dawid's conditions, resulting in the weak robustification seen in Section A.1. One way to produce the desired robustification (in the sense of De Finetti 1961) and still retain elements of the KF's recursive mechanism is to assume *independent* Student-$t$ distributions for all of the (now presumed scalar) variables in question. Doing so, we not only weaken one of Dawid's conditions; we also obtain an almost sure convergence of pertinent densities. This is summarized in the following theorem.

*Theorem 1.* Let $g(x)$, the prior density at $x$, be a Student-$t$ density with $n$ degrees of freedom (df), and let $f(y - x)$, the density at $Y$ given $x$, also be a Student-$t$ density with location $x$ and $m$ df. Then if $n > m$, $h(x \mid y)$, the posterior density at $x$ given $y$, will converge a. e. to $g(x)$ as $|y| \to \infty$. Furthermore, the posterior distribution will converge to the prior distribution uniformly over Borel sets.

*Proof.* Suppose that $X \sim t(x; 0, 1, n)$ and $(Y \mid x) \sim t(y; x, s, m)$, where $t(\cdot, x, s, m)$ denotes a Student-$t$ distribution with location $x$, scale $s$, and df $m$, and let $n > m$. Due to symmetry, it suffices to consider the case $y \to +\infty$. Since $t(y; x, s, m) = t(x; y, s, m)$,

$$h(x \mid y) = \frac{[1 + x^2/n]^{-(n+1)/2}[1 + (x - y)^2/ms]^{-(m+1)/2}}{\int_{-\infty}^{\infty} [1 + z^2/n]^{-(n+1)/2}[1 + (z - y)^2/ms]^{-(m+1)/2}\, dz}.$$

To show that the posterior distribution of $X$ will converge uniformly to its prior we argue that $h(x \mid y)$ converges to $g(x)$ a. e. with respect to a Lebesgue measure $\lambda$ on $(\mathcal{R}, \mathcal{B})$. We then use Scheffe's theorem (Billingsley 1968, p. 224) to complete the proof. To prove a. e. convergence, note that for any $x \in \mathcal{R}$, the ratio

$$g(x)/h(x/y) = [1 + (x - y)^2/ms]^{(m+1)/2} \times \int_{-\infty}^{\infty} t(z; 0, 1, n)[1 + (z - y)^2/ms]^{-(m+1)/2}\, dz$$

can be made arbitrarily close to unity for sufficiently large $y$. The argument is based on the observation that for any fixed $x$ and large $y$, $g(x)/h(x \mid y)$ can be represented as the sum of integrals, one of which becomes as close to unity (and the others as close to 0) as we wish. The algebraic details are in Meinhold (1984).

*Theorem 2.* Under the conditions of Theorem 1, if $m > n$, $h(x \mid y)$ will converge a. e. to the likelihood $(y \mid x)$ as $|y| \to \infty$; the convergence is uniform over Borel sets.

*Proof.* The proof is analogous to that of Theorem 1, except that now, for any $x \in \mathcal{R}$, the ratio

$$f(y - x)/h(x \mid y) = (1 + x^2/n)^{(n+1)/2} \int_{-\infty}^{\infty} t(z; y, s, m)[1 + z^2/n]^{-(n+1)/2}\, dz$$

can be made arbitrarily close to unity for sufficiently large $y$.

Note that Theorems 1 and 2 are duals of each other. When $n = m$, a form of "outlier confusion" would occur in the sense that $h(x \mid y)$ converges to neither the prior nor the likelihood.

## 4. IMPLEMENTATION OF THEOREM 1 IN THE KF SETTING

In KF models, provision is made for both the observable $Y_t$ and the unobservable $\theta_t$ to experience spurious disturbances and distortions. In this article, however, we concentrate on "outliers" in the observation equation only; these are referred to as *additive outliers* (see Martin and Raftery 1987). Accordingly, we assume that $\theta_0, w_1, \ldots, w_T, v_1, \ldots, v_T$ are independent, each having a Student-$t$ density, with $\theta_0, w_1, \ldots, w_T$ having $n$ df and $v_1, \ldots, v_T$ having $m$ df with $m < n$. The joint density of all the variables is not a multivariate Student-$t$ density (as described in Sec. A.1), and the resulting posterior distributions will have "poly-$t$" densities (Broemeling 1985, p. 447) with no closed-form representation. Therefore, we need to propose a recursive approximation scheme to circumvent this difficulty yet adhere to the conditions of Theorem 1. Prior to developing the approximation scheme, however, we must recognize that Theorem 1 describes only the behavior of the posterior distribution in the face of an extreme or limiting observation. We also need to outline the nature of the posterior distribution in the presence of a "moderate"observation.

### 4.1 Behavior of the Posterior in the Presence of a Moderate Observation

Suppose that the prior distribution of $\theta_t$ and the likelihood centered at $Y_t$ are both Student-$t$ densities, with the prior having a greater number of df than the likelihood. The difference between the location of these two densities is $e_t = Y_t - F_t G_t \hat{\theta}_{t-1}$. When $e_t$ is small, the posterior distribution of $\theta_t$ is unimodal; as $e_t$ increases (a decrease yielding mirror-image results) the mode of the posterior distribution begins to move to the right, but at a decreasing rate. Eventually, the rightward shift of the single mode ceases and a second mode emerges to the right of the first. As $e_t$ continues to increase, the left-side mode reverses direction and shifts toward $G_t \hat{\theta}_{t-1}$, the mode of the prior, whereas the right-side mode follows the rightward movement of $Y_t$. This phenomenon of the appearance of the second mode and its divergence from the first is similar to that encountered by O'Hagan (1981), who aptly termed it a "moment of indecision." Ultimately, as $e_t$ becomes very large, the result of Theorem 1 comes into play and the leftmost mode of the posterior distribution essentially coincides with the mode of the prior, and the probability mass under the rightmost mode becomes essentially negligible—resulting in a posterior distribution that is practically indistinguishable from the prior. In what follows, we propose a plan of approximation that mimics this behavior.

### 4.2 Approximating the Posterior Distribution

The scheme proposed here is motivated by Jeffreys (1961, sec. 4.2), and it begins with the idea that the posterior distribution of $\theta_{t-1}$ may be represented by a mixture of Student-$t$ densities. The prior for $\theta_t$ is then formed by a componentwise convolution of each member of the mixture and $w_t$. This produces a prior with the same number of components as the posterior, each with unchanged measure of location (the distribution of $w_t$ being centered at 0) but with new dispersion. When an observation $Y_t$ arises, Bayes's theorem is applied componentwise to the prior mixture; for each member, a posterior distribution results, which may be unimodal or bimodal. In the first instance, a single Student-$t$ density is used to approximate this updated component; in the second, a mixture of two $t$ densities is used. Specifically, suppose that after time $t - 1$, the posterior distribution of $\theta_{t-1}$ is represented by a mixture of $N_{t-1}$ Student-$t$ densities, each with $n$ df; that is,

$$\theta_{t-1} \sim \sum_{j=1}^{N_{t-1}} \alpha_{j,t-1} t(\cdot; \mu_{j,t-1}, \sigma^2_{j,t-1}, n),$$

with $\sum_{j=1}^{N_{t-1}} \alpha_{j,t-1} = 1$.

Because of (1.2) and prompted by a cumulant matching scheme suggested by Patil (1965), the prior for $\theta_t$ will be approximated by the mixture

$$\theta_t \sim \sum_{j=1}^{N_{t-1}} \alpha_{j,t-1} t(\cdot; G_t \mu_{j,t-1}, G_t^2 \sigma^2_{j,t-1} + W_t, n).$$

Simplicity and convenience are the key virtues of this ap-

proximation; its suitability should be judged in the light of its performance, such as in the example in Section 4.5. Upon receipt of $Y_t$, the likelihood will be described by $t(\cdot; y_t, V_t, m)$, where $m < n$. A prior-to-posterior analysis peformed for each component of the mixture that constitutes the prior gives $N_{t-1}$ posteriors, each given via Bayes's law as

$$\frac{t(x; G_t\mu_{j,t-1}, G_t^2\sigma_{j,t-1}^2 + W_t, n)t(x; y_t, V_t, m)}{\int_{-\infty}^{\infty} t(z; G_t\mu_{j,t-1}, G_t^2\sigma_{j,t-1}^2 + W_t, n)t(z; y_t, V_t, m)\, dz}. \tag{4.1}$$

Each of the $N_{t-1}$ posteriors will be either unimodal or bimodal, depending on the nature of the roots of a cubic equation (see Sec. A.2). When one of these posteriors is unimodal, it will be approximated by a Student-$t$ density with $n$ df and $\mu_{j,t}$ set equal to (say) $x_0$, the mode of the posterior. The weight $\alpha_{j,t}$ assigned to this component in the mixture remains unchanged and is equal to $\alpha_{j,t-1}$. The scale $\sigma_{j,t}^2$ of the approximating Student-$t$ density is determined by setting

$$t(x_0; \mu_{j,t} = x_0, \sigma_{j,t}^2, n) = \Gamma\left(\frac{n+1}{2}\right) \Big/ \left[\Gamma\left(\frac{n}{2}\right)(\pi n\sigma_{j,t}^2)^{1/2}\right],$$

the *height* of the approximating density at $x_0$, equal to the height of the actual posterior (4.1) at $x_0$.

When one of the individual posteriors is bimodal, it will be approximated by a mixture of two Student-$t$ densities, each with $n$ df. These densities will be centered at $\mu_{j1,t}$ set equal to (say) $x_1$, the smaller mode of the posterior, and $\mu_{j2,t}$ set equal to (say) $x_2$, the larger mode of the posterior. The scale parameters $\sigma_{j1,t}^2$ and $\sigma_{j2,t}^2$ of the two mixing Student-$t$ densities will be found by equating the *curvature* of the approximating densities at their modes $x_1$ and $x_2$ to the (approximate) curvature of the actual posterior (4.1) at $x_1$ and $x_2$, respectively. Specifically, to obtain $\sigma_{j1,t}^2$ we would set

$$-\frac{d^2}{dx^2}\log t(x; \mu_{j1,t} = x_1, \sigma_{j1,t}^2, n)\,|_{x=x_1} = \frac{n+1}{n\sigma_{j1,t}^2}$$

so that

$$-\frac{d^2}{dx^2}\log t(x; G_t\mu_{j,t-1}, G_t^2\sigma_{j,t-1}^2 + W_t, n)|_{x=x_1}$$

$$= \frac{n+1}{nG_t^2\sigma_{j,t-1}^2}\left\{(x_1 - G_t\mu_{j,t-1})^2\left[\frac{1 + (x_1 - G_t\mu_{j,t-1})^2}{n(G_t^2\sigma_{j,t-1}^2 + W_t)}\right]^{-2} - [1 + (x_1 - G_t\mu_{j,t-1})^2/n(G_t^2\sigma_{j,t-1}^2 + W_t)]^{-1}\right\}.$$

Implicit in this calculation is the assumption ($w$ log) that $Y_t$ lies to the right of $G_t\mu_{j,t-1}$ and that, accordingly, this latter quantity is closer to $x_1$ and $Y_t$ is relatively far from $x_1$ (a necessary condition for the bimodal case to arise). So the curvature of $t(x; y_t, V_t, m)$ is negligible in the neighborhood of $x_1$. By an analogous argument, $\sigma_{2j,t}^2$ is determined by setting

$$-\frac{d^2}{dx^2}\log t(x; x_2, \sigma_{2j,t}^2, n)|_{x=x_2} = -\frac{d^2}{dx^2}\log t(x; Y_t, V_t, m).$$

The weights $\alpha_{j1,t}$ and $\alpha_{j2,t}$ of the mixture components will be given as

$$\alpha_{j1,t} = \alpha_{j,t-1}t(G_t\mu_{j,t-1}; Y_t, V_t, m)/[t(G_t\mu_{j,t-1}; Y_t, V_t, m) + t(Y_t; G_t\mu_{j,t-1}, G_t^2\sigma_{j,t-1}^2 + W_t, n)]$$

$$\alpha_{j2,t} = \alpha_{j,t-1} - \alpha_{j1,t}. \tag{4.2}$$

The approximation scheme could be refined—for example, by considering the exact curvature of the actual posterior instead of the approximate curvature or by using a mixture to approximate a posterior that is unimodal but skewed. Such fine-tuning may, in some applications, be preferable to the simple approach selected here.

### 4.3 Choice of Degrees of Freedom $n$ and $m$

The extent to which our proposed scheme accepts or rejects an observation depends on the values selected for $n$ and $m$. Requiring that $n > m > 0$ restricts the choice to values in the north-northeast octant. For illustration we consider three cases: Case A, in which $n = 30$ and $m = 29$; Case B, in which $n = 30$ and $m = 2$; and Case C, in which $n = 2$ and $m = 1$. Case A reflects the fact that both the prior and the likelihood are near normal; Case B reflects the fact that the prior is near normal, whereas the likelihood is very heavy-tailed; Case C reflects the fact that both the prior and the likelihood are heavy-tailed.

In Figure 1 we show, via box-and-whisker plots, the posterior distributions for each of the three cases, for a range of values of $Y_t$; specifically, $Y_t = 0, 1, 2, 3, 4, 5, 10$, and 20. The horizontal widths of the boxes represent the $\alpha$ weight given to that component of the posterior mixture; these weights are determined via (4.2). The right end of each of the three illustrations shows the box-and-whisker plots of the prior distributions; these would correspond to the posterior distributions when $Y_t$ is infinite.

An examination of the plots in Figure 1 shows that for Case A, wherein the information provided by the likelihood is slightly "less strong" than that from the prior, the "rejection" of $Y_t$ and the return of the posterior to the prior is slow, requiring large values of $Y_t$. In Case B, the much lighter-tailed prior dominates the likelihood (so that little weight is given to $Y_t$) and the posterior returns to the prior for small values of $Y_t$. In Case C, there is a scant amount of information in both the prior and the likelihood, so the indecision (multimodality) in the posterior arises quickly and continues to persist, even for large values of $Y_t$.

The disparate kinds of behavior of the posterior distributions illustrated in Figure 1 may serve as a guide to users in choosing values of $n$ and $m$ so as to attain a desired degree of robustness.

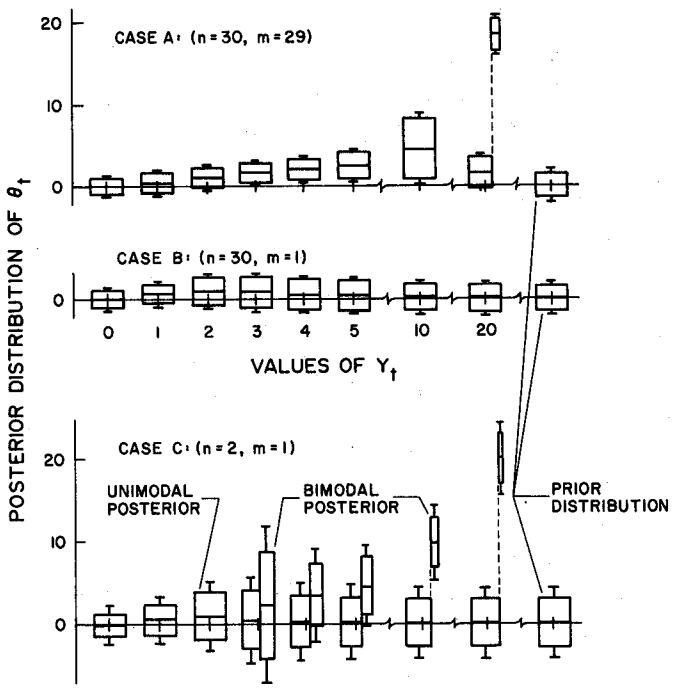


*Figure 1. Box-and-Whisker Plots Illustrating the Behavior of the Posterior of $\theta_t$ for Different Values of $Y_t$, n, and m. The width assigned to each box in a plot of the bimodal posterior indicates the $\alpha$ weight assigned to that component of the mixture.*

## 4.4 Some Comments on the Approximation Scheme

An aspect of the approximation scheme deserving of comment is the growing number of components in the mixture that approximates the posterior distribution of the state of nature. A similar situation has been encountered in other work (e.g., Harrison and Stevens 1976). A strategy for curtailing this growth is to merge two adjacent mixture components that are close enough to yield a unimodal density when combined. A sufficient condition that suggests the desirability of such a combination is that $|\mu_{j,t} - \mu_{k,t}| < \min(\sigma_{j,t}, \sigma_{k,t})$, for any $j \neq k$. The combined (pooled) component is assigned location $\mu^*_{j,t} = \alpha_{j,t}\mu_{j,t} + \alpha_{k,t}\mu_{k,t}$, scale $\sigma^{*2}_{j,t} = \alpha_{j,t}\sigma^2_{j,t} + \alpha_{k,t}\sigma^2_{k,t} + \alpha_{j,t}\alpha_{k,t}(\mu_{j,t} - \mu_{k,t})^2$, and weight $\alpha^*_{j,t} = \alpha_{j,t} + \alpha_{k,t}$. Another possibility is that some components of the mixture come about as a result of outlying values of $Y_t$ that may not recur. In such cases, the $\alpha$ weights associated with such components gradually diminish, and eventually we may choose simply to drop such components from the mixture, when in our judgment (say $\alpha = .01$ or $.05$) they become of little significane. Some caution should be exercised against discarding mixture elements too quickly, however, or the model's ability to respond to genuine shifts in structure is compromised—that is, those cases where a deviant $Y_t$ results from an "outlier" in the system equation error, the effect of which will persist, as opposed to one from the observation equation. For such dramatic shifts in the model structure, the posterior mass in the area of extreme observations will, as Theorem 1 suggests, be quite small. Nonetheless, should similar values of $Y_t$ consistently continue to be realized, then under our proposed scheme for component weights, one-half of the mass of the posterior will "relocate" to the appropriate vicinity, providing a signal to the modeler about a change in the structure of the process and indicating that the original model specification is no longer realistic. This may be viewed as a desirable property for a recursive tracking procedure to possess—especially an automated one. Robust filters with data-dependent mean squared error recursions, in the style of Masreliez (1975), also have the desirable relocation or "regaining of tracking" property (see Martin and Thompson 1982; Martin and Yohai 1985).

Finally, our use of the Student-$t$ distribution essentially results in a bounded-influence filter with a smoothly redescending influence, which leads to a rejection-at-infinity property. From a practical point of view, there is little difference between a very small weight and total rejection, so the approach here can be regarded as a mechanism that, in practice, does not operate substantially differently from other robust procedures, but does pursue the goal of robustness (as enunciated by Box 1980), entirely within the strictures of the Bayesian paradigm (as interpreted by De Finetti 1961).

## 4.5 An Illustrative Example

We illustrate the operation of the proposed scheme by considering a simple simulated example described in Table 1 and Figure 2. For each time period $t$ ($t = 1, \ldots, 10$) we assume a structure known as the *steady model,* wherein $Y_t = \theta_t + v_t$; $\theta_t = \theta_{t-1} + w_t$; and (in the notation of Sec. 4.2) $v_t \sim t(\cdot; 0, 2, 2)$, $w_t \sim t(\cdot; 0, 1, 3)$, and $\theta_0 \sim t(\cdot; 0, 1, 3)$. The likelihood is then approximated by $t(\cdot; Y_t, 2, 2)$, and the prior distribution for $\theta_t$ is given by $\sum_{j=1}^{N_{t-1}} \alpha_{j,t-1} t(\cdot;$

*Table 1. A Simulated Example*

| $t$ | $Y_t$ | $j$ | $\alpha_j$ | $\mu_{j,t}$ | $\sigma^2_{j,t}$ |
|---|---|---|---|---|---|
| 0 | | 1 | 1 | 0 | 1 |
| 1 | 1.638 | 1 | 1 | .856 | 1.031 |
| 2 | −.224 | 1 | 1 | .283 | .878 |
| 3 | 9.540 | 1 | .698 | .759 | 2.117 |
| | | 2 | .302 | 8.908 | 2.688 |
| 4 | −2.079 | 1 | .698 | −1.123 | 1.771 |
| | | 2 | .146 | −1.590 | 2.390 |
| | | 3 | .156 | 8.071 | 4.451 |
| 4* | | 1 | .844 | −1.204 | 2.231 |
| | | 2 | .156 | 8.071 | 4.451 |
| 5 | .181 | 1 | .844 | −.303 | 1.119 |
| | | 2 | .156 | .793 | 5.305 |
| 6 | 1.902 | 1 | .844 | .877 | 1.307 |
| | | 2 | .156 | 1.663 | 1.300 |
| 6* | | 1 | 1 | 1.000 | 1.390 |
| 7 | 6.721 | 1 | .539 | 2.189 | 4.266 |
| | | 2 | .461 | 5.555 | 5.437 |
| 8 | 8.134 | 1 | .539 | 7.400 | 3.884 |
| | | 2 | .461 | 7.635 | 1.639 |
| 8* | | 1 | 1 | 7.508 | 2.863 |
| 9 | 5.882 | 1 | 1 | 6.379 | 1.241 |
| 10 | 7.003 | 1 | 1 | 6.726 | .840 |

* A condensation of the original.

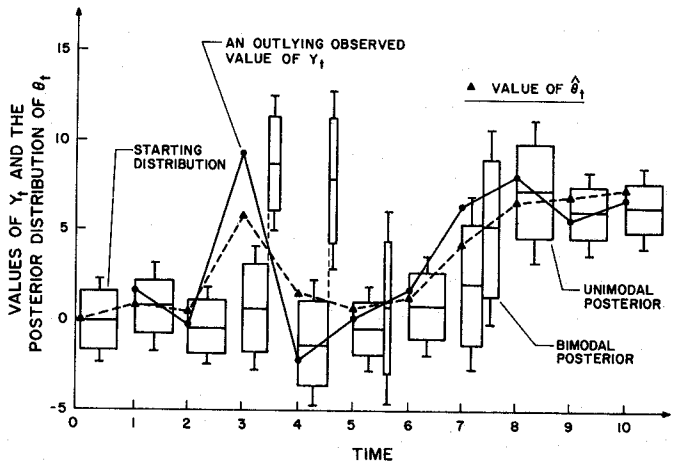


*Figure 2. Box-and-Whisker Plot of the Posterior Distribution of $\theta_t$ for the Data in Table 1. The width assigned to each box in a plot of the bimodal posterior indicates the $\alpha$ weight assigned to that component of the mixture. Also shown are values of $\hat{\theta}_t$, given by the nonrobust filter.*

$\mu_{j,t-1}, \sigma^2_{j,t-1} + 1, 3)$. The resultant posterior distributions are displayed in Figure 2, via box-and-whisker plots, with the horizontal width of the box representing the $\alpha$ weight given to that component of the posterior mixture. In those instances wherein the displayed posterior distribution represents a condensation of the original (i.e., some components of the original mixture have been combined; see Sec. 4.4), in Table 1 an asterisk is indicated on the $t$ associated with it.

Two outlier situations are considered. At $t = 3$ a large value of $Y_t$ occurs, but the subsequent values of $Y_t$ return to the earlier neighborhood of 0; note that by the time $t = 6$, the posterior distribution of $\theta_t$ is "back to normal." At $t = 7$ a permanent shift in $Y_t$ occurs corresponding to a dramatic shift in the state of nature; we see from Figure 2 that the ensuing sequence of posteriors responds in a reasonable manner to this shift. Contrast this with the behavior of a nonrobust Kalman filter for which the values $\hat{\theta}_t$, obtained via (1.3), are for $t = 1, \ldots, 10$, given as .819, .298, 4.919, 1.420, .800, 1.351, 4.036, 6.085, 5.984, and 6.493, respectively. Finally, from an inspection of the plots of Figure 1, it appears that if one desires to minimize the occurrence of bimodal posteriors, yet require a fast response to a permanent shift in the level of the process, then one must opt for large values of $m$ and $n$, with $n > m$.

## 5. SUMMARY AND CONCLUSIONS

We have proposed a model wherein, by using the Bayesian approach with a judiciously selected family of density functions, belief regarding an unknown state of nature (subsequent to observation of a related random variable) is embodied in a mixture of probability density functions from the specified family. A system for updating this expression of knowledge in the face of succeeding observations is suggested, along with some ideas on how a modeler might restrain the growth of this probabilistic representation, depending on his or her objectives. Finally, according to the restraints imposed, one can produce a mechanism yielding expression of belief about the state of nature that will be largely unaffected by occasional spurious observations (bounded and smoothly redescending influence), yet will eventually respond to genuine but dramatic shifts in this unknown quantity (relocation or regaining of tracking property). Our proposed approach easily lends itself for implementation on a computer. We give an example illustrating the manner in which it responds to spurious observations and changes in the state of nature.

## APPENDIX: CONSIDERATION OF THE MULTIVARIATE–$T$ DISTRIBUTION AND OTHER DERIVATIONS

### A.1 The KF Model With Multivariate Student-$T$ Distributed Variables

Suppose that $\theta_t$ and $Y_t$ are of dimension $p \times 1$ and $q \times 1$, respectively, and that $t = 1, \ldots, T$. Suppose that the $T(p + q) + p$ dimension column vector $[\theta_0, v_1, w_1, \ldots, v_t, w_t]'$ is assumed to have a multivariate Student-$t$ distribution with $\upsilon$ df, location $\boldsymbol{\mu} = [\hat{\theta}_0, 0, \ldots, 0]$, and scale $J_{T(p+q)+p}$, where the latter is a $[T(p + q) + p] \times [T(p + q) + p]$ matrix with block-diagonal elements $\Sigma_0, V_1, W_1, \ldots, V_T, W_T$, and off-diagonal elements 0. For convenience, this distribution, which has a density at $\mathbf{x}$ given by

$$f_{t_p}(\mathbf{x}; \boldsymbol{\mu}, J, \upsilon) = \frac{\Gamma\left(\frac{\upsilon + p}{2}\right)|J|^{-1/2}}{\Gamma\left(\frac{\upsilon}{2}\right)(\pi\upsilon)^{p/2}}[1 + (\mathbf{x} - \boldsymbol{\mu})'J^{-1}(\mathbf{x} - \boldsymbol{\mu})/\upsilon]^{-[(\upsilon+p)/2]},$$

is denoted by $t_{T(p+q)+p}(\cdot; \boldsymbol{\mu}, J, \upsilon)$. Then we can easily prove the following theorem (details are omitted).

*Theorem A.1.* For a multivariate-$t$–derived KF, $\theta_t \sim t_p(\cdot; \hat{\theta}_t, C(Y_t, \ldots, Y_1)\Sigma_t, \upsilon + qt)$, where $\hat{\theta}_t$ and $\Sigma_t$ are given by (1.3) and (1.4), respectively, and $C(Y_t, \ldots, Y_1) = [\upsilon + \sum_{s=1}^{t}(Y_s - F_sG_s\hat{\theta}_{s-1})'(F_sR_sF_s' + V_s)^{-1}(Y_s - F_sG_s\hat{\theta}_{s-1})]/(\upsilon + qt)$.

From Theorem A.1 it is clear that a multivariate-$t$–dervied KF possesses a closure property that enables us to derive an inference mechanism via an exact distribution theory; however, it still yields a nonrobust measure of location. On the other hand, the scale parameter is a multiple of $\Sigma_t$ (the Gaussian-derived KF scale parameter), with $C(Y_t, \ldots, Y_1)$ as the multiplier. The magnitude of the multiplier increases quadratically in the one-step-ahead prediction error $|Y_s - F_sG_s\hat{\theta}_{s-1}|$, implying that observations that are inconsistent with prior expectations cause a loss in precision of the estimate of the state of nature. Thus this filter neither accepts nor rejects a large observation in the sense that even though the measure of location is a linear function of an outlying $Y_t$, the measure of dispersion explodes. The posterior, in effect, tends toward a non-informative distribution; we call this behavior *outlier confusion.* Since inference is concerned with the entire posterior distribution (not just the measure of location), the use of a multivariate-$t$–derived KF may be considered a weak robustification of the KF model.

Another feature of the multivariate-$t$–derived KF emanates from the fact that $C(Y_t, \ldots, Y_1)$ possesses a memory of unlimited order. In the absence of extreme outliers and assuming that $\Sigma_0$, the $V_t$'s, and the $W_t$'s were realistically valued, $C(Y_t, \ldots, Y_1)$ converges in probability (as $t \to \infty$) to 1. Hence the long-term tendency for $C(Y_t, \ldots, Y_1)$ to adhere to or deviate from

1 provides information about the reasonableness of the model specification. Finally, since $\theta_t$ has a posterior density with $v + tq$ df, the growing information of the process means that $\theta_t$ behaves in a less heavy-tailed manner as compared with $Y_t$, which always has $v$ df.

### A.2 Roots of a Cubic Equation Characterizing the Posterior

Suppose that the prior density of $X$ is a Student-$t$ density, denoted by $t(x; 0, 1, n)$, and that the likelihood of $Y$ given $X = x$ is $t(y; x, s, m) = t(x; y, s, m)$, $n > m$. Because of symmetry we need only consider the case $y > 0$. The posterior density of $X \mid Y = y$ is

$$h(x \mid y) \propto [1 + x^2/n]^{-(n+1)/2}[1 + (x - y)^2/ms]^{-(m+1)/2}. \quad \text{(A.1)}$$

The mode(s) of the posterior density will be located at those values of $x$ for which $d\{\log[h(x \mid y)]\}/dx = 0$, where

$$\begin{aligned} d\{\log[h(x \mid y)]\}/dx \propto & -(n+1)x[ms + (x-y)^2] \\ & -(m+1)(x-y)[n + x^2] \\ = & -[(m+1) + (n+1)]x^3 \\ & + [(m+1) + 2(n+1)]yx^2 \\ & -[(m+1)n + (n+1)(ms + y^2)]x \\ & + (m+1)ny. \end{aligned} \quad \text{(A.2)}$$

Methods for solving this cubic equation may be found in Burington (1955). To cast (A.2) into the "standard" form, $x^3 + px^2 + qx + r = 0$, we let

$$\begin{aligned} k &= (n+1)/(m+1) \\ p &= -[(2k+1)/(k+1)]y \\ q &= [n + k(ms + y^2)]/(k+1) \\ r &= -ny/(k+1). \end{aligned} \quad \text{(A.3)}$$

Setting $z = x + p/3$, we arrive at the reduced form $z^3 + az + b = 0$, where

$$\begin{aligned} a &= \frac{3q - p^2}{3} = \frac{-(k^2 + k + 1)y^2 + 3(k+1)(n + kms)}{3(k+1)^2} \\ b &= (2p^3 - 9pq + 27r)/27 \\ &= (2k+1)(k+2)(k-1)y^3 + 9(k+1)[(2k+1) \\ &\quad \times (n + kms) - 3n(k+1)]y. \end{aligned} \quad \text{(A.4)}$$

For $d = b^2/4 + a^3/27$, the following will hold: (a) If $d > 0$ there is one real root and two complex conjugate roots, implying that the posterior is unimodal. (b) If $d \leq 0$ there are three real roots (all unequal if $d < 0$), implying that the posterior is bimodal.

It is easy to demonstrate that (A.2) is positive for $x \leq 0$ but negative for $x \geq y$, so all roots lie in the open interval $(0, y)$. Therefore, a necessary condition for bimodality is $a < 0$; from (A.4), this requires that $y > [3(k+1)(n + kms)/(k^2 + k + 1)]^{1/2}$, so the posterior will be unimodal whenever $y$ is sufficiently small. On the other hand, $d$ may be seen to be a sixth-degree equation in $y$, with a negative coefficient of $y^6$, so $d$ will be negative and the posterior will be bimodal for all $y$ greater than some constant.

By expressing the three real roots in the bimodal situation as $-\text{sign}(b)2(-a/3)^{1/2}\cos(\phi + j\pi/3)$, $j = 0, 1, 2$, where $\phi$ is obtained as a solution to $\cos(3\phi) = [-(b^2/4) \div (a^3/27)]^{1/2}$, it may be shown that as $y$ becomes very large, the two modes will be located near 0 and $y$. The height of the posterior at the rightmost mode will approach 0 relative to the height at the leftmost mode. A first-order Taylor series approximation to the leftmost mode—the "dominant" one that returns to the prior mode of 0—locates it near $ny/[n + k(ms + y)^2]$, which exhibits the property of an initial increase and a subsequent decrease of the location of the mode.

[*Received February 1987. Revised September 1988.*]

## REFERENCES


Anderson, B. D., and Moore, J. B. (1979), *Optimal Filtering,* Englewood Cliffs, NJ: Prentice-Hall.

Berger, J. O. (1984), "The Robust Bayesian Viewpoint," in *Robustness of Bayesian Analysis,* ed. J. Kadane, Amsterdam: Elsevier, pp. 63–144.

Billingsley, P. (1968), *Convergence of Probability Measures,* New York: John Wiley.

Boncelet, C. G., Jr., and Dickinson, B. W. (1984), "A Variant of Huber Robust Regression," *SIAM Journal on Scientific and Statistical Computing,* 5, 720–734.

Box, G. E. P. (1980), "Sampling and Bayes' Inference in Scientific Modeling and Robustness" (with discussion), *Journal of the Royal Statistical Society,* Ser. A, 143, 383–430.

Broemeling, L. (1985), *Bayesian Analysis of Linear Models,* New York: Marcel Dekker.

Burington, R. (1955), *Handbook of Mathematical Tables and Formulas,* Sandusky, OH: Handbook Publishers.

Dawid, A. P. (1973), "Posterior Expectations for Large Observations," *Biometrika,* 60, 664–667.

De Finetti, B. (1961), "The Bayesian Approach to the Rejection of Outliers," in *Proceedings of the Fourth Berkeley Symposium on Probability and Statistics,* Berkeley: University of California Press, pp. 199–210.

Diderrich, G. T. (1985), "The Kalman Filter From the Perspective of Goldberger–Theil Estimators," *The American Statistican,* 39, 193–198.

Ershov, A. A., and Lipster, R. S. (1978), "Robust Kalman Filter in Discrete Time," *Automation and Remote Control,* 39, 359–367.

Gather, U., and Kale, B. K. (1986), "Outlier Generating Models—A Review," Technical Report 117, Iowa State University, Dept. of Statistics.

Green, R. F. (1974), "A Note on Outlier-Prone Families of Distribution," *The Annals of Statistics,* 2, 1293–1295.

——— (1976), "Outlier-Prone and Outlier-Resistant Distributions," *Journal of the American Statistical Association,* 71, 502–505.

Guttman, I., and Pena, D. (1985), "Robust Filtering," Comment on "Dynamic Generalized Linear Models and Bayesian Forecasting," by M. West, P. J. Harrison, and H. S. Migon, *Journal of the American Statistical Association,* 80, 91–92.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics,* New York: John Wiley.

Harrison, P. J., and Stevens, C. F. (1976), "Bayesian Forecasting" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 38, 205–247.

Hill, B. M. (1974), "On Coherence, Inadmissibility and Inference About Many Parameters in the Theory of Least Squares," in *Studies in Bayesian Econometrics and Statistics,* eds. S. E. Fienberg and A. Zellner, Amsterdam: North-Holland, pp. 555–587.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford, U.K.: Clarendon Press.

Kalman, R. E. (1960), "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering,* 82, 34–45.

Kitagawa, G. (1987), "Non-Gaussian State–Space Modeling of Nonstationary Time Series" (with discussion), *Journal of the American Statistical Association,* 82, 1032–1063.

Leonard, T. (1974), "A Modification to the Bayes Estimate for the Mean of a Normal Distribution," *Biometrika,* 61, 627–628.

Lindley, D. V. (1968), "The Choice of Variables in Multiple Regression" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 30, 31–66.

Martin, R. D. (1979), "Approximate Conditional-Mean Type Smoothers and Interpolators," in *Smoothing Techniques for Curve Estimation,* eds. T. Gesser and M. Rosenblatt, Berlin: Springer-Verlag, pp. 117–143.

Martin, R. D., and Raftery, A. E. (1987), "Robustness, Computation, and Non-Euclidean Models," *Journal of the American Statistical Association,* 82, 1044–1050.

Martin, R. D., and Thompson, D. J. (1982), "Robust Resistant Spectrum Estimation," *IEEE Proceedings,* 70, 1097–1115.

Martin, R. D., and Yohai, V. J. (1985), "Robustness in Time Series and Estimating ARMA Models," in *Handbook of Statistics 5,* eds. E. J.

Hannan, P. R. Krishnaiah, and M. M. Rao, New York: Elsevier, pp. 119–155.
Masreliez, C. J. (1975), "Approximate Non-Gaussian Filtering With Linear State and Observation Relations," *IEEE Transactions on Automatic Control,* 20, 107–110.
Masreliez, C. J., and Martin, R. D. (1975), "Robust Estimation via Stochastic Approximation," *IEEE Transactions on Information Theory,* 21, 263–271.
——— (1977), "Robust Bayesian Estimation for Linear Model and Robustifying the Kalman Filter," *IEEE Transactions on Automatic Control,* 22, 361–371.
Meeden, G., and Isaacson, D. (1977), "Approximate Behavior of the Posterior Distribution for a Large Observation," *The Annals of Statistics,* 5, 899–908.
Meinhold, R. J. (1984), "A Robustification of the Kalman Filter via the Bayesian Approach," unpublished doctoral dissertation, George Washington University, Graduate School of Arts and Sciences.
Meinhold, R. J., and Singpurwalla, N. D. (1983), "Understanding the Kalman Filter," *The American Statistician,* 37, 123–127.
——— (1987), "A Kalman Filter Smoothing Approach for Extrapolations in Certain Dose–Response, Damage-Assessment, and Accelerated-Life-Testing Studies," *The American Statistician,* 41, 101–106.
Morris, J. M. (1976), "The Kalman Filter: A Robust Estimator for Some Classes of Linear Quadratic Problems," *IEEE Transactions on Information Theory,* 22, 526–534.
Neyman, J., and Scott, E. L. (1971), "Outlier Proneness of Phenomena and of Related Distributions," in *Optimizing Methods in Staistics,* ed. J. S. Rustagi, New York: Academic Press, pp. 413–430.
O'Hagan, A. (1981), "A Moment of Indecision," *Biometrika,* 68, 329–330.
——— (1987), "Modelling With Heavy Tails," in *Bayesian Statistics 3,* eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Oxford, U.K.: Clarendon Press, pp. 345–359.
Patil, V. H. (1965), "Approximations to the Behrens–Fisher Distribution," *Biometrika,* 52, 267–271.
Smith, A. F. M., and West, M. (1983), "Monitoring Renal Transplants: An Application of the Multiprocess Kalman Filter," *Biometrics,* 39, 867–878.
Sorensen, H. W., and Alspach, D. L. (1971), "Recursive Bayesian Estimation Using Gaussian Sums," *Automatica,* 7, 467–479.
Tsai, C., and Kurz, L. (1983), "An Adaptive Robustizing Approach to Kalman Filtering," *Automatica,* 19, 279–288.
West, M. (1981), "Robust Sequential Approximate Bayesian Estimation," *Journal of the Royal Statistical Society,* Ser. B, 43, 157–166.
——— (1982), "Aspects of Recursive Bayesian Estimation," unpublished Ph.D. dissertation, University of Nottingham (U.K.), Dept. of Mathematics.
West, M., Harrison, P. J., and Migon, H. S. (1985), "Dynamic Generalized Linear Models and Bayesian Forecasting," *Journal of the American Statistical Association,* 80, 73–97.
Wiener, N. (1949), *Extrapolation, Interpolation and Smoothing of Stationary Time Series,* Cambridge, MA: MIT Press.
Zellner, A. (1976), "Bayesian and Non-Bayesian Analysis of the Regression Model With Multivariate Student-*t* Error Terms," *Journal of the American Statistical Association,* 71, 400–405.