Finite-Memory Denoising in Impulsive Noise Using Gaussian Mixture Models

Yonina C. Eldar, *Student Member, IEEE,* and Arie Yeredor, *Member, IEEE*

***Abstract*—We propose an efficiently structured nonlinear finite-memory filter for denoising (filtering) a Gaussian signal contaminated by additive impulsive colored noise. The noise is modeled as a zero-mean Gaussian mixture (ZMGM) process. We first derive the optimal estimator for the static case, in which a Gaussian random variable (RV) is contaminated by an impulsive ZMGM RV. We provide an analytical derivation of the resulting mean-squared error (MSE), and compare the performance to that of the optimal linear estimator, identifying cases of significant improvement. Building upon these results, we develop a suboptimal finite-memory filter for the dynamic case, which is nearly optimal in the minimum MSE sense. The resulting filter is a nonlinearly weighted combination of a fixed number of linear filters, for which a computationally efficient architecture is proposed. Substantial improvement in performance over the optimal linear filter is demonstrated using simulation results.**

I. Introduction

A COMMON approach to modeling background noise in various applications is to use a Gaussian model for the noise distribution. In estimating (filtering, denoising) a stochastic process contaminated by additive noise, when both the signal and noise distributions are modeled as Gaussian, the optimal filter in the mean-squared error (MSE) sense is a linear filter, e.g., taking the form of Wiener or Kalman filters. However, in many physical environments, the noise exhibits impulsive characteristics, which cannot be adequately described by a Gaussian model.

Several approaches exist for accommodating impulsive noise in the context of filtering. A rather naive but often efficient approach is to use preprocessing hard limiters, thus, practically discarding the outliers. However, when either the signal or noise involved have a substantial correlation length, such a hard-limiting operation may be far from optimal.

Other, more elaborate approaches (e.g., [6], [4], [5], and [1]), use explicit non-Gaussian statistical models to describe the impulsive behavior of the noise. One such model, which gained increasing popularity in the past decade, is the alpha-stable model, used, e.g., in [5] and [1]. A drawback of this model is the relative complexity of both the analytical derivations and filter implementations involved. Another possible model for describing impulsive noise is the Gaussian mixture (GM) model, used, e.g., in [6] and [4].

GM modeling is popular in the signal-processing community mainly in the context of speech recognition. However, little research effort has been directed at GM modeling of time series. Such modeling warrants the use of standard linear models, with the traditional Gaussian driving noise substituted by a GM noise, e.g., to describe impulsively driven autoregressive (AR) processes [11].

In this paper we propose to model the impulsive, possibly colored noise as a linearly filtered sequence of independent, identically distributed (iid) GM random variables (RVs) consisting of zero-mean components, which we refer to as zero-mean GM (ZMGM). Such modeling is appealing in several respects: ZMGM encompasses the popular zero-mean Gaussian model as a special case; the sum of ZMGM and/or Gaussian RVs is also a ZMGM RV; when a ZMGM process undergoes linear filtering, the output is also a ZMGM process. In addition, a ZMGM model is most appropriate for describing outlier situations, by assuming mixtures of two components, where the first component occurs with a high probability, and the second component has a significantly larger variance and occurs with a small probability.

Modeling background noise as a GM iid sequence has been studied in depth by Sorenson and Alspach [6]. They derived the optimal MSE estimator for a Gaussian signal contaminated by GM noise, and pursued a recursive implementation thereof. Their estimator consists of a bank of Kalman filters, whose outputs are combined with proper weighting. However, a severe drawback of this optimal approach is that, as the number of GM components involved grows exponentially in time, so does the number of filters in the bank, rendering the estimator computationally impractical. In [4], Masreliez proposed a suboptimal approximation to the Sorenson and Alspach recursive filter, which alleviates the computational load by deliberately ignoring the non-Gaussian characteristic of some intermediate conditional distributions along the way.

We propose an alternative, nonrecursive approach to the filtering problem, in which the estimation of the desired signal at each time-instant is based only on a fixed number of most recent observations. The use of finite memory inherently limits the exponential increase of the number of ZMGM components involved. Moreover, by employing a certain approximation, many of these components can be justly ignored. Thus, the resulting finite-memory filter serves as a close approximation of the optimal finite-memory filter, but is yet computationally appealing. By exploiting some algebraic properties of the proposed filter, we proceed to derive an efficient implementation

Manuscript received August 9, 2000; revised October. This paper was recommended by Associate Editor O. Tanrikulu.

Y. C. Eldar was with the Department of Electrical Engineering–Systems, Tel-Aviv University, Tel-Aviv 69978, Israel. She is now with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: yonina@mit.edu).

A. Yeredor is with the Department of Electrical Engineering–Systems, Tel-Aviv University, Tel-Aviv 69978, Israel (e-mail: arie@eng.tau.ac.il).

Publisher Item Identifier S 1057-7130(01)11309-1.

structure, which further reduces the overall computational complexity.

As in [6] and [4], it is assumed throughout that all the statistical model parameters of both the desired signal and the noise are known.

The paper is structured as follows: In Section II, we confine the discussion to the case in which both the signal and noise are iid time series, where the optimal filtering is equivalent to optimal memoryless (scalar) RV estimation. We derive the optimal MSE estimator and provide (in the Appendix) an analytical derivation for evaluating the resulting MSE, which we use to identify cases of interest. These results are then extended in Section III, where we develop the nearly optimal fixed-memory filter for estimating a stationary Gaussian signal contaminated by additive colored ZMGM noise. In Section IV, we provide some simulations results demonstrating the MSE improvement over the optimal linear filter. Concluding remarks are in Section V.

Throughout the paper, $x \sim \mathcal{N}(\mu, \sigma^2)$ denotes an RV x taking a Gaussian distribution with mean μ and variance σ^2. We denote vectors by boldface lowercase letters, and matrices by boldface uppercase letters. All RVs involved are real valued.

II. Optimal Estimation of a Gaussian Random Variable in ZMGM Noise

We first consider the problem in which it is desired to estimate a RV $x \sim \mathcal{N}(0, \sigma_x^2)$ from the noisy measurement $y = x + v$, where v is some impulsive noise, statistically independent of x. We model the noise v as a GM with M zero-mean components (i.e., ZMGM) of variances $\sigma_1^2 < \sigma_2^2 < \ldots < \sigma_M^2$ occurring with probabilities $p_1, p_2, \ldots p_M$, respectively. The noise probability distribution function (pdf) is given by

$$f_v(v) = \frac{1}{\sqrt{2\pi}} \sum_{m=1}^{M} \frac{p_m}{\sigma_m} e^{-\frac{v^2}{2\sigma_m^2}} \tag{1}$$

where $\sum_{m=1}^{M} p_m = 1$.

The pdf of (1) can be given the following interpretation. Let i denote an auxiliary RV which we call a "component indicator," that takes on values from 1 to M, and indicates the Gaussian component from which v is drawn. Then, $P(i = m) = p_m$ and $f_{v|i}(v \,|\, i = m)$ is $\mathcal{N}(0, \sigma_m^2)$ for $m = 1, 2, \ldots M$. Consequently, $f_v(v)$ can be expressed as $\sum_{m=1}^{M} f_{v|i}(v \,|\, i = m) P(i = m)$.

Since x is zero-mean Gaussian, the measurement $y = x + v$ is a ZMGM RV, taking variances

$$s_m^2 \triangleq \sigma_x^2 + \sigma_m^2, \qquad m = 1, 2, \ldots M \tag{2}$$

with probabilities $p_m, m = 1, 2, \ldots, M$, respectively.

The optimal minimum MSE (MMSE) estimator $\hat{x}(y)$ of x is well known to be the conditional expectation $\hat{x}(y) = E[x \,|\, y]$. To evaluate $\hat{x}(y)$, we rewrite this expectation as

$$\begin{aligned} E[x \,|\, y] &= E[E[x \,|\, y, i = m]] = E[\hat{x}_m(y)] \\ &= \sum_{m=1}^{M} q_m(y) \hat{x}_m(y) \end{aligned} \tag{3}$$

where $\hat{x}_m(y) = E[x \,|\, y, i = m]$ and $q_m(y) = P(i = m \,|\, y)$ is the posterior probability of $i = m$ given y, and is equal to

$$\begin{aligned} q_m(y) &= \frac{f_{y|i}(y \,|\, i = m) P(i = m)}{f_y(y)} \\ &= \frac{\frac{p_m}{s_m} \exp\left(-\frac{y^2}{2s_m^2}\right)}{\sum_{k=1}^{M} \frac{p_k}{s_k} \exp\left(-\frac{y^2}{2s_k^2}\right)}. \end{aligned} \tag{4}$$

Noting that given $i = m$, x and y are jointly Gaussian, $\hat{x}_m(y)$ is the optimal linear estimator of x assuming that $i = m$, or equivalently, that $v \sim \mathcal{N}(0, \sigma_m^2)$. Thus,

$$\hat{x}_m(y) = \frac{\sigma_x^2}{s_m^2} y \triangleq h_m y \tag{5}$$

where $h_m = \sigma_x^2 / s_m^2$. Substituting (5) and (4) into (3), the MMSE linear estimator is given by

$$\hat{x}(y) = \sum_{m=1}^{M} q_m(y) \hat{x}_m(y) = \frac{\sum_{m=1}^{M} \frac{p_m}{s_m} \exp\left(-\frac{y^2}{2s_m^2}\right) h_m}{\sum_{k=1}^{M} \frac{p_k}{s_k} \exp\left(-\frac{y^2}{2s_k^2}\right)} y. \tag{6}$$

This estimator can be viewed as a weighted combination of the (conditional) optimal linear estimators $\hat{x}_1, \hat{x}_2, \ldots \hat{x}_M$ defined in (5). However, the weighting coefficients introduce nonlinearity in the measurement y. To visualize the effect of the nonlinearity we compare the optimal estimator to the optimal linear estimator $\hat{x}_L(y)$ of x from $y = x + v$, which is given by

$$\hat{x}_L(y) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2} y \tag{7}$$

where $\sigma_v^2 = \sum_{m=1}^{M} p_m \sigma_m^2$ denotes the variance of the noise v.

Fig. 1 demonstrates the behavior of the two estimators superimposed on 500 realizations of true values x versus measurements y for the case of $M = 2$ with $p_1 = 0.9, \sigma_1 = 1, p_2 = 0.1, \sigma_2 = 10$, and a signal-to-noise ratio (SNR) of $\sigma_x^2 / \sigma_v^2 = 0.5$. From the figure, it is seen that the linear estimator attempts to compensate for occurrences of outliers by using a moderate slope, which in turn misaligns with the more probable population in the center. The optimal estimator, on the other hand, uses a weighted combination of two linear estimators $\hat{x}_1(y)$ and $\hat{x}_2(y)$, such that the slope of each is properly aligned with the population appearing in its region of domination. For example, the region of domination of $\hat{x}_2(y)$ is the region of large values of y, where $q_2(y) \approx 1$. The resulting optimal estimator is reminiscent of a hard limiter, which is a popular tool for estimation in the presence of impulsive noise. However, while a hard limiter completely discards suspected outliers, the optimal estimator gradually decreases its sensitivity as the posterior probability of an outlier increases.

In the Appendix we derive an expression for computing the MSE attained by the optimal estimator $\hat{x}(y)$ for the case of two mixture components ($M = 2$). We used this expression to demonstrate in Fig. 2 the relative improvement in MSE attained by the optimal estimator over the optimal linear estimator, whose MSE is given by $\sigma_x^2 \sigma_v^2 / (\sigma_x^2 + \sigma_v^2)$. We consider three values of SNR: low (-3 dB), medium (0 dB), and high

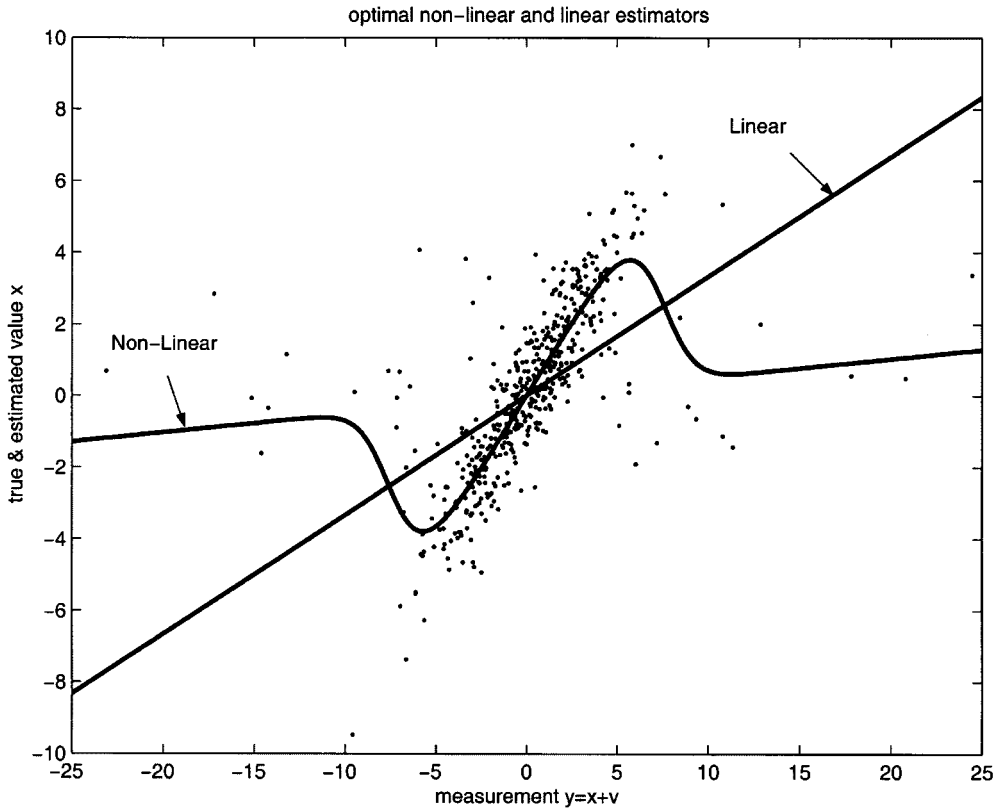

Fig. 1. Optimal nonlinear and linear estimators of x from $y = x + v$, superimposed on 500 realizations.

(7 dB). For each SNR, we present six curves corresponding to $\sigma_2/\sigma_1 = 2, 5, 10, 15, 20, and 25$, showing the relative improvement as a function of the outlier probability p_2. It is evident that the improvement increases as σ_2/σ_1 increases, but attains an optimum as a function of p_2. The peaks become sharper and lower as the SNR increases.

It is interesting to note that in the inverse situation, when it is desired to estimate the noise component, the optimal estimator is the complementary estimator, i.e., $\hat{v}(y) = y - \hat{x}(y)$. This holds true since if $y = x + v$, then $E[v \mid y] = E[y - x \mid y] = y - E[x \mid y]$. Since $\hat{v}(y) - v = -(\hat{x}(y) - x)$, the resulting MSEs are the same in both cases. This property is also shared by the optimal linear estimator so that the relative improvement remains the same. This observation is useful when analyzing the case of an impulsive (ZMGM) iid time series contaminated by iid Gaussian noise.

III. Estimating Gaussian Signals in Colored ZMGM Noise

We now expand the results of the previous section to include estimation of correlated time series contaminated by impulsive colored noise.

Let $x[n]$ denote the desired signal, which is stationary zero-mean Gaussian with correlation function $R_{xx}[l]$. It is desired to filter (estimate) $x[n]$ from the noisy measurements $y[n] = x[n] + v[n]$, where $v[n]$ denotes the noise, modeled as a correlated ZMGM process, independent of $x[n]$. The correlated ZMGM noise is assumed to be an order K moving-average [MA(K)] process, i.e., it is created by an iid sequence $w[n]$ of ZMGM RVs passing through a finite-impulse response (FIR) filter of length $K+1$, with coefficients $g[0], g[1], \ldots g[K]$. For simplicity, we assume that $w[n]$ is ZMGM with $M = 2$ mixture components of variances $\sigma_1^2 < \sigma_2^2$ appearing with probabilities p_1 and p_2, respectively; the results extend in a straightforward way to arbitrary values of M. For notational brevity, we denote by p the outlier probability, thus, $p_2 = p, p_1 = 1 - p$.

To derive a computationally appealing estimator, we restrict the discussion to finite-memory filters of length L. Thus, it is desired to filter $x[n]$ from the preceding L observations of $y[n]$, i.e., from the vector $\boldsymbol{y}_n = [y[n] y[n-1] \cdots y[n-L+1]]^T$.

From Fig. 2, it is evident that in the case of memoryless estimation when $v[n]$ is ZMGM with $M = 2$, the improvement attained by using optimal estimation over optimal linear estimation is substantial for small values of p. We, therefore, focus our discussion on such cases in the context of finite-memory filtering as well.

The derivation of our suboptimal filter is based on the further assumption, that the probability of multiple occurrences of outliers in $w[n]$ influencing $\boldsymbol{y}_n$ is negligible. Since $\boldsymbol{y}_n$ contains L samples, and $w[n]$ is filtered by a filter of length $K+1$, there are $K+L$ samples of $w[n]$ bearing impact on $\boldsymbol{y}_n$. The probability of no outlier occurrence in these $K+L$ samples is given by $P_0 = (1-p)^{K+L}$, and the probability of occurrence of a single outlier is given by $P_1 = (K+L)p(1-p)^{K+L-1}$. Therefore, the probability of multiple occurrences of outliers in $w[n]$ influencing $\boldsymbol{y}_n$ is given by $1 - P_0 - P_1$, which when expanded ignoring terms

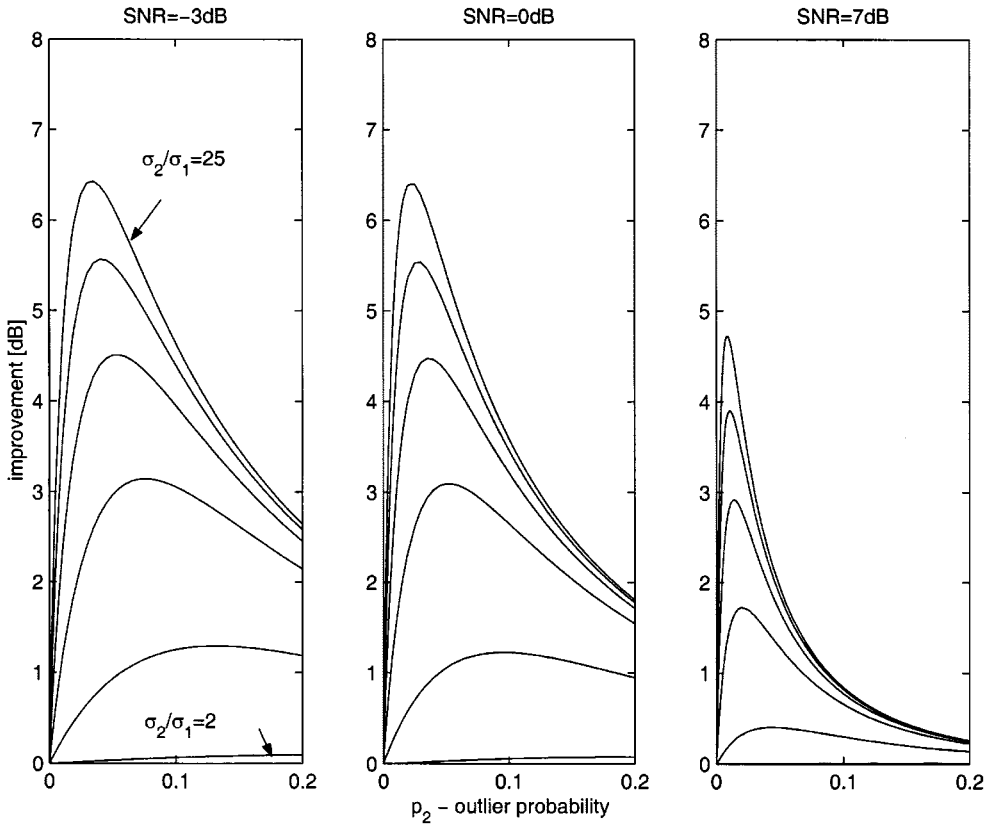

Fig. 2. MSE improvement [dB] attained by the nonlinear ZMGM estimator over the optimal linear estimator for three SNRs: low, medium, and high. Curves in each figure show the improvement as a function of the outlier probability p_2, parameterized by $\sigma_2/\sigma_1 = 2, 5, 10, 15, 20, 25$ (the curves are monotonically ordered in σ_2/σ_1, with the highest corresponding to $\sigma_2/\sigma_1 = 25$).

smaller than p^2 reduces to $p^2(K+L)(K+L-1)/2$. Our assumption can, therefore, be expressed by the condition

$$[(K+L)p]^2 \ll 1. \tag{8}$$

Thus, rather than develop the optimal finite-memory estimator, we develop a suboptimal finite-memory estimator, which is approximately optimal under the highly probable assumption (8) that no more than one outlier occurred in the recent $K+L$ samples of $w[n]$.

To derive our suboptimal estimator of $x[n]$, we define a vector $\boldsymbol{i}_n = [i_n\, i_{n-1}\, \dots\, i_{n-(K+L-1)}]^T$ composed of the individual "component indicators" for $w[n], w[n-1], \dots w[n-(K+L-1)]$, respectively [as described in Section II in the paragraph following (1))].

In analogy to (3), the optimal MSE estimator $\hat{x}[n]$ of $x[n]$ is given by the conditional expectation $E[x[n]\,|\,\boldsymbol{y}_n] = E[E[x[n]\,|\,\boldsymbol{y}_n, \boldsymbol{i}_n]]$, where $\boldsymbol{i}_n$ can take 2^{K+L} values. However, for our suboptimal estimator we consider only the $K+L+1$ most probable values of $\boldsymbol{i}_n$ corresponding to one or no outlier occurrence, namely $\boldsymbol{i}^0, \boldsymbol{i}^1, \dots \boldsymbol{i}^{K+L}$, where $\boldsymbol{i}^0 = [1\ 1\ \dots\ 1]^T$ and $\boldsymbol{i}^k,\ k = 1, 2, \dots K+L$ are all-ones vectors with a 2 at the kth entry (the remaining possibilities will be indexed implicitly as $\boldsymbol{i}^k, k = K+L+1, \dots 2^{K+L}-1$).

Thus, in a way similar to the scalar case (6), our estimator can be expressed as

$$\hat{x}[n] = \sum_{k=0}^{K+L} q_k(\boldsymbol{y}_n)\hat{x}_k[n] \tag{9}$$

where $\hat{x}_k[n] = E[x[n]\,|\,\boldsymbol{y}_n, \boldsymbol{i}_n = \boldsymbol{i}^k]$, and $q_k(\boldsymbol{y}_n) = P(\boldsymbol{i}_n = \boldsymbol{i}^k\,|\,\boldsymbol{y}_n)$ denotes the posterior probability of $\boldsymbol{i}_n = \boldsymbol{i}^k$ given $\boldsymbol{y}_n$.

To determine $\hat{x}_k[n]$ we note that given $\boldsymbol{i}_n = \boldsymbol{i}^k, x[n]$ and $\boldsymbol{y}_n$ are jointly Gaussian. Thus,

$$\hat{x}_k[n] = E[x[n]\,|\,\boldsymbol{y}_n, \boldsymbol{i}_n = \boldsymbol{i}^k] = \boldsymbol{R}_{xy}(\boldsymbol{R}_{yy}^k)^{-1} \tag{10}$$

where $\boldsymbol{R}_{yy}^k = \boldsymbol{R}_{xx} + \boldsymbol{R}_{vv}^k$ and $\boldsymbol{R}_{xy} = \boldsymbol{r}_{xx}^T = [R_{xx}[0] R_{xx}[1] \dots R_{xx}[L-1]]$. Here, $\boldsymbol{R}_{xx}$ is the $L \times L$ correlation matrix of $x[n], \boldsymbol{r}_{xx}$ is its first column, and $\boldsymbol{R}_{vv}^k$ is the conditional correlation matrix of $v[n]$ given $\boldsymbol{i}_n = \boldsymbol{i}^k$, which can be derived as follows:
denote by

$$\boldsymbol{G} = \begin{bmatrix} g[0] & g[1] & \cdots & g[K] & 0 & \cdots & 0 \\ 0 & g[0] & \cdots & & g[K] & \ddots & \vdots \\ \vdots & \ddots & & & & & 0 \\ 0 & \cdots & 0 & g[0] & g[1] & \cdots & g[K] \end{bmatrix} \tag{11}$$

the $L \times (K+L)$ Toeplitz matrix composed of the noise-generating FIR filter's impulse response coefficients, $g[l],\ l = 0, 1, \dots K$. We now have

$$\boldsymbol{v}_n = \boldsymbol{G}w_n \tag{12}$$

where $\boldsymbol{v}_n \triangleq [v[n] v[n-1] \dots v[n-L+1]]^T$ and $\boldsymbol{w}_n \triangleq [w[n] w[n-1] \dots w[n-K-L+1]]^T$. Thus,

$$\boldsymbol{R}_{vv}^k = \boldsymbol{G}\boldsymbol{R}_{ww}^k\boldsymbol{G}^T \tag{13}$$

where $\boldsymbol{R}_{ww}^{k}$ is the correlation matrix of $\boldsymbol{w}_n$ given $\boldsymbol{i}_n = \boldsymbol{i}^k$, derived as follows: for $k = 0$ there are no outliers in $\boldsymbol{w}_n$, and, therefore, all the components have equal variance σ_1^2. For $k \neq 0$, there is a single outlier of variance σ_2^2 at the kth entry of $\boldsymbol{w}_n$. In any event, the components of $\boldsymbol{w}_n$ are independent. Thus,

$$\boldsymbol{R}_{ww}^{k} = \begin{cases} \sigma_1^2 \boldsymbol{I}, & k = 0 \\ \sigma_1^2 \boldsymbol{I} + \delta^2 \boldsymbol{e}_k \boldsymbol{e}_k^T, & k \neq 0 \end{cases} \tag{14}$$

where $\boldsymbol{I}$ denotes the identity matrix, $\boldsymbol{e}_k$ is the kth column of $\boldsymbol{I}$, and $\delta^2 = \sigma_2^2 - \sigma_1^2$. Therefore,

$$\boldsymbol{R}_{vv}^{k} = \begin{cases} \sigma_1^2 \boldsymbol{G}\boldsymbol{G}^T, & k = 0 \\ \sigma_1^2 \boldsymbol{G}\boldsymbol{G}^T + \delta^2 \boldsymbol{g}_k \boldsymbol{g}_k^T, & k \neq 0 \end{cases} \tag{15}$$

where $\boldsymbol{g}_k$ is the kth column of $\boldsymbol{G}$.
Noting that

$$\boldsymbol{R}_{yy}^{k} = \boldsymbol{R}_{yy}^{0} + \delta^2 \boldsymbol{g}_k \boldsymbol{g}_k^T, \qquad k = 1, 2, \ldots K+L \tag{16}$$

where

$$\boldsymbol{R}_{yy}^{0} = \boldsymbol{R}_{xx} + \sigma_1^2 \boldsymbol{G}\boldsymbol{G}^T \tag{17}$$

and using the Matrix Inversion Lemma (see, e.g., [2, p. 18]) we have

$$\left(\boldsymbol{R}_{yy}^{k}\right)^{-1} = \left(\boldsymbol{R}_{yy}^{0}\right)^{-1} - \frac{\delta^2}{1 + \delta^2 \boldsymbol{g}_k^T \left(\boldsymbol{R}_{yy}^{0}\right)^{-1} \boldsymbol{g}_k} \times \left[\left(\boldsymbol{R}_{yy}^{0}\right)^{-1} \boldsymbol{g}_k\right] \left[\left(\boldsymbol{R}_{yy}^{0}\right)^{-1} \boldsymbol{g}_k\right]^T \tag{18}$$

and

$$\det\left\{\boldsymbol{R}_{yy}^{k}\right\} = \det\left\{\boldsymbol{R}_{yy}^{0}\right\} \left(1 + \delta^2 \boldsymbol{g}_k^T \left(\boldsymbol{R}_{yy}^{0}\right)^{-1} \boldsymbol{g}_k\right). \tag{19}$$

Finally,

$$q_k(\boldsymbol{y}_n) = \frac{f_{\boldsymbol{y}|\boldsymbol{i}}(\boldsymbol{y}_n \,|\, \boldsymbol{i}_n = \boldsymbol{i}^k) P(\boldsymbol{i}_n = \boldsymbol{i}^k)}{\sum_{k=0}^{2^{K+L}-1} f_{\boldsymbol{y}|\boldsymbol{i}}(\boldsymbol{y}_n \,|\, \boldsymbol{i}_n = \boldsymbol{i}^k) P(\boldsymbol{i}_n = \boldsymbol{i}^k)} \approx \frac{f_{\boldsymbol{y}|\boldsymbol{i}}(\boldsymbol{y}_n \,|\, \boldsymbol{i}_n = \boldsymbol{i}^k) P(\boldsymbol{i}_n = \boldsymbol{i}^k)}{\sum_{k=0}^{K+L} f_{\boldsymbol{y}|\boldsymbol{i}}(\boldsymbol{y}_n \,|\, \boldsymbol{i}_n = \boldsymbol{i}^k) P(\boldsymbol{i}_n = \boldsymbol{i}^k)} \tag{20}$$

where

$$P(\boldsymbol{i}_n = \boldsymbol{i}^k) = \begin{cases} (1-p)^{K+L}, & k = 0 \\ p(1-p)^{K+L-1}, & k = 1, 2, \ldots K+L \end{cases} \tag{21}$$

and

$$f_{\boldsymbol{y}|\boldsymbol{i}}(\boldsymbol{y}_n \,|\, \boldsymbol{i}_n = \boldsymbol{i}^k) = \frac{1}{\sqrt{\det\left\{2\pi \boldsymbol{R}_{yy}^{k}\right\}}} \times \exp\left(-\frac{1}{2} \boldsymbol{y}_n^T \left(\boldsymbol{R}_{yy}^{k}\right)^{-1} \boldsymbol{y}_n\right). \tag{22}$$

(The $\approx$ in (20) should be interpreted in a probabilistic sense: the terms discarded in the denominator are negligible only in the highly probable case where not more than one outlier occurred in $\boldsymbol{w}_n$.)
Using (18) and (19), we may rewrite (22) as

$$\begin{aligned} f_{\boldsymbol{y}|\boldsymbol{i}}(\boldsymbol{y}_n \,|\, \boldsymbol{i}_n = \boldsymbol{i}^k) &= \frac{1}{\sqrt{\det\left\{2\pi R_{yy}^{0}\right\}}} \times \exp\left(-\frac{1}{2} \boldsymbol{y}_n^T \left(R_{yy}^{0}\right)^{-1} \boldsymbol{y}_n\right) \cdot \frac{1}{\sqrt{1 + \delta^2 \boldsymbol{g}_k^T \left(\boldsymbol{R}_{yy}^{0}\right)^{-1} \boldsymbol{g}_k}} \\ &\quad \times \exp\left(\frac{1}{2} \frac{\delta^2}{1 + \delta^2 \boldsymbol{g}_k^T \left(\boldsymbol{R}_{yy}^{0}\right)^{-1} \boldsymbol{g}_k} \left(\boldsymbol{y}_n^T \left(\boldsymbol{R}_{yy}^{0}\right)^{-1} \boldsymbol{g}_k\right)^2\right). \end{aligned} \tag{23}$$

Note that the first two terms are independent of k and are, therefore, common to the numerator and denominator in (20), which we may consequently rewrite as

$$q_k(\boldsymbol{y}_n) = \frac{\eta_k(\boldsymbol{y}_n)}{\sum_{k=0}^{K+L} \eta_k(\boldsymbol{y}_n)} \tag{24}$$

where

$$\eta_k(\boldsymbol{y}_n) = \begin{cases} 1 - p, & k = 0 \\ \gamma_k \exp\left(\beta_k \left(\tilde{\boldsymbol{h}}_k^T \boldsymbol{y}_n\right)^2\right), & k \neq 0 \end{cases} \tag{25}$$

with

$$\tilde{\boldsymbol{h}}_k = \left(\boldsymbol{R}_{yy}^{0}\right)^{-1} \boldsymbol{g}_k \tag{26a}$$

$$\gamma_k = p / \sqrt{1 + \delta^2 \boldsymbol{g}_k^T \tilde{\boldsymbol{h}}_k} \tag{26b}$$

$$\beta_k = \delta^2 \gamma_k^2 / 2p^2. \tag{26c}$$

Substituting $q_k(\boldsymbol{y}_n)$ into (9), we observe that $\hat{x}[n]$ can be interpreted as a weighted combination of linear estimators. The first linear estimator, $\hat{x}_0[n]$, is the optimal linear estimator of $x[n]$ assuming that $\boldsymbol{v}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_{vv}^{0})$, namely, that no outlier occurred in $\boldsymbol{w}_n$. The other $K+L$ estimators are each optimal linear estimators assuming that $\boldsymbol{v}_n \sim \mathcal{N}(\boldsymbol{u}_o, \boldsymbol{R}_{vv}^{k})$, namely that a single outlier occurred in the respective location in $\boldsymbol{w}_n$. The weights $q_k(\boldsymbol{y}_n)$ reflect the posterior probability of the respective events presumed by the estimators.

The data-dependent terms $\tilde{\boldsymbol{h}}_k^T \boldsymbol{y}_n$ in the estimator are common to both the linear estimators $\hat{x}_k[n]$ and to the terms $\eta_k(\boldsymbol{y}_n)$ for $k = 1, 2, \ldots K+L$, respectively. This feature can be exploited by the implementation depicted in Fig. 3, where $\boldsymbol{h}_0 = (\boldsymbol{R}_{yy}^{0})^{-1} \boldsymbol{r}_{xx}, \alpha_k = 2\beta_k \boldsymbol{r}_{xx}^T \tilde{\boldsymbol{h}}_k$, and $\tilde{\boldsymbol{h}}_k, \gamma_k$ and β_k are given above in (26). The estimator is calculated using only $K+L+1$ FIR filters whose outputs are denoted $\hat{x}_0[n], z_1[n], \ldots z_{K+L}[n]$, scalar nonlinearities composed of simple $(\cdot)^2$ and $\exp(\cdot)$ operations, and one division. This architecture also allows parallel implementation of the FIR filters and nonlinear operations (up to the final division).

It is important to observe that the filter's memory length L has to be carefully designed. On one hand, the correlation length

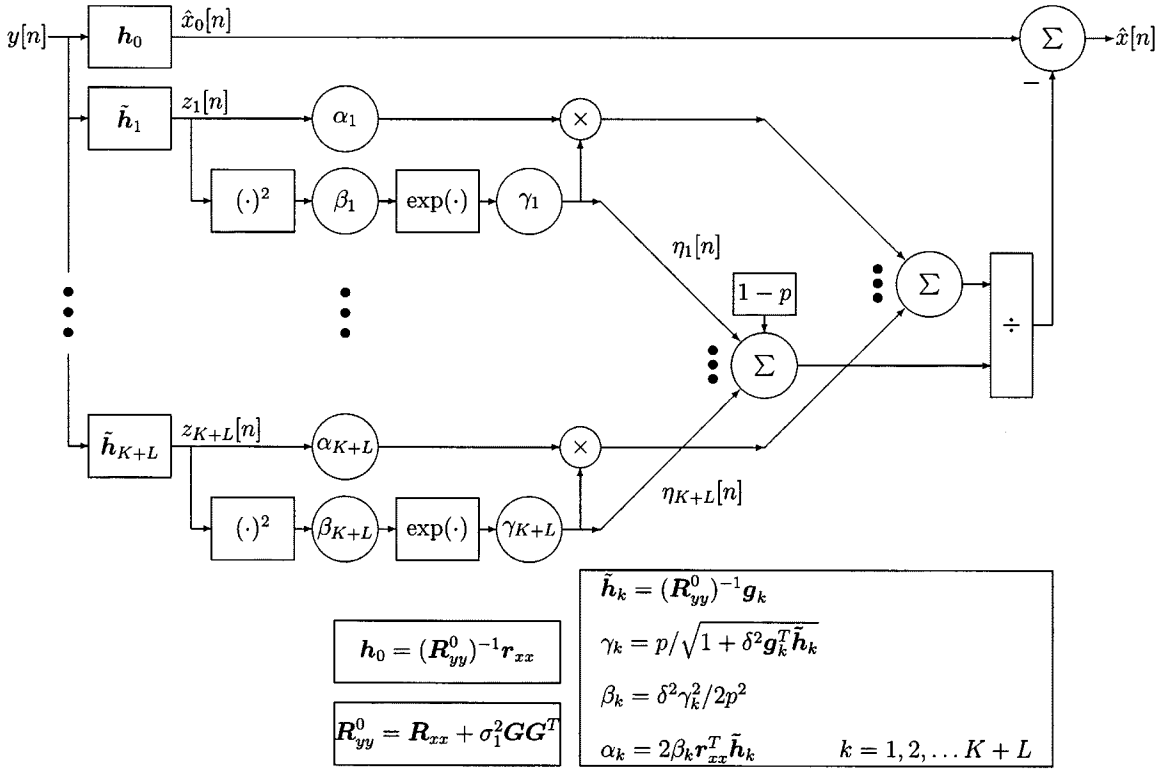

Fig. 3. Finite memory filter for Gaussian signals in colored ZMGM noise.

of the signals involved may require long memory; on the other hand, as L grows, the condition $[(K+L)p]^2 \ll 1$ of (8) may be breached, implying that more than one outlier is likely to occur in $\boldsymbol{w}_n$, thus, devalidating our filter derivation. It has been verified by simulation (some of which are presented in the next section), that when L is chosen properly, substantial improvement of MSE over the optimal linear estimator can be attained by the proposed filter. The improvement increases as σ_2/σ_1 increases, but attains an optimum as a function of SNR and p.

As we noted in the scalar case, when it is desired to estimate the noise component, the optimal estimator is the complementary estimator, i.e., $\hat{v}[n] = y[n] - \hat{x}[n]$. This holds true since if $y[n] = x[n] + v[n]$, then $E[v[n]\,|\,\boldsymbol{y}_n] = E[y[n] - x[n]\,|\,\boldsymbol{y}_n] = y[n] - E[x[n]\,|\,\boldsymbol{y}_n]$. Since $\hat{v}[n] - v[n] = -(\hat{x}[n] - x[n])$, the resulting MSEs are the same in both cases. However, since our estimator is only nearly optimal, this property is only approximately exhibited.

IV. Simulations Results

We present some simulation results demonstrating the improvement attained by the proposed filter (denoted hereafter as the "ZMGM filter") over the optimal linear filter. The underlying signal was generated as an AR process of order 2, with poles at $z = 0.95$ and $z = 0.9$, i.e., it satisfies the difference equation

$$x[n] = 1.85x[n-1] - 0.855x[n-2] + u[n] \qquad (27)$$

where $u[n]$ is a zero-mean white Gaussian noise sequence, whose variance was set such that $x[n]$ would have unit variance. The additive colored ZMGM noise was generated by passing an iid ZMGM sequence $w[n]$ through a five-coefficients FIR filter $g[l]$

$$g[l] = \begin{cases} l+1, & 0 \le l \le 4 \\ 0, & \text{otherwise} \end{cases} \qquad (28)$$

hence,

$$v[n] = w[n] + 2w[n-1] + 3w[n-2] + 4w[n-3] + 5w[n-4]. \qquad (29)$$

The ZMGM sequence $w[n]$ consists of $M = 2$ zero-mean Gaussian components with variances σ_1 and σ_2 occurring with probabilities $1 - p$ and p, respectively—thus, p denotes the "outlier probability." σ_1 and σ_2 were set such that $v[n]$ would have unit variance, and their ratio σ_2/σ_1 equals its desired values (see the following).

The measured signal is $y[n] = x[n] + v[n]$, hence, the input SNR to all filters tested is 0 dB. To estimate $x[n]$ from $y[n]$, we applied $y[n]$ to three filters: the optimal linear causal ("infinite memory") Kalman filter, the optimal linear FIR filter of length L, and the proposed ZMGM nonlinear filter of memory length L. Results are displayed in terms of the SNR at the filters' outputs, which is actually also the MSE in estimating $x[n]$, since $x[n]$ has unit variance.

In Fig. 4 we demonstrate the dependence of performance on the ratio σ_2/σ_1, with the outlier probability fixed at $p = 0.02$. For the finite-memory filters (FIR and ZMGM), we used a memory length of $L = 6$. The solid lines represent the theoretical output SNRs for the Kalman and FIR filters, as indicated. These values are, of course, independent of the variances ratio.

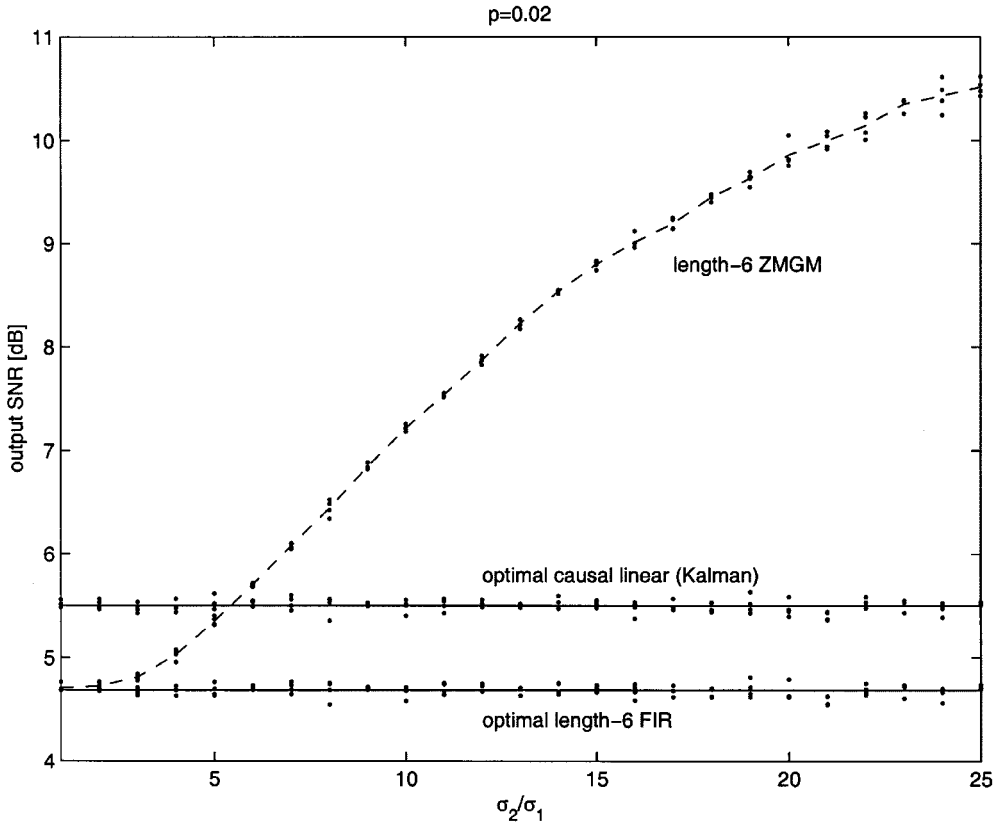

Fig. 4. SNR at the output of the optimal linear filters and proposed ZMGM filter versus σ_2/σ_1. Solid lines denote the theoretical values for the optimal linear Kalman (infinite memory) filter and an FIR filter of length $L = 6$. Simulations results are indicated by dots, each representing a single trial over a signal of length 250 000 (four trials per test point). All filters used the same data.

Superimposed on these lines are dots representing simulation results [four trials (dots) per test point]. Similar dots are also provided for the ZMGM filter as indicated, where their empirical means are connected by a dashed line (we do not have a tractable analytic expression for the ZMGM filter's performance—the derivation in the Appendix is only valid for the static case). It is clearly seen that the ZMGM filter offers substantial improvement over the optimal linear FIR filter of the same memory length, as well as over the infinite-memory Kalman filter, as the variance ratio σ_2/σ_1 increases. As expected, for $\sigma_2/\sigma_1 = 1$ (practically no outlier situation), the ZMGM filter's performance coincides with that of the FIR filter, but it quickly departs as the outliers situation becomes more imminent.

In Fig. 5, we demonstrate the interesting dependence on the memory length L. In this case, the variances ratio is fixed at $\sigma_2/\sigma_1 = 25$, and we present results for two outlier probabilities: $p = 0.02$ and $p = 0.05$. Again, solid lines represent the theoretical values, and simulation results appear as dots, with four trials (dots) per test point. Only the linear FIR and the ZMGM filters depend on L; the Kalman filter's results are simply repeated for each L for ease of comparison. As expected, the linear FIR filter's SNR approaches the Kalman filter's SNR from below as L increases. However, both fall well below the SNR of the ZMGM filter. It is reassuring to note, however, that while the FIR filter's performance monotonically improves as L is increased, the ZMGM filter's performance has an optimum as a function of L, since further increase of L gradually breeches the condition in (8), and thus, slowly degrades performance. This behavior is more pronounced for the higher outlier probability $(p = 0.05)$. In addition, since the ZMGM performance has an optimum as a function of p (see, e.g., Fig. 1), performance for $p = 0.05$ is generally worse than for $p = 0.02$. Naturally, the performance of the linear filters is insensitive to p.

V. Conclusion

We presented the optimal (nonlinear) estimator and associated error analysis for estimating a Gaussian RV from its measurement contaminated by impulsive noise modeled as a ZMGM RV. The estimator can be interpreted as a nonlinear weighting of optimal linear estimators, each suited to a corresponding mixture component. The weighting reflects the posterior probability of occurrence of the respective components. In extreme outlier situations the optimal estimator resembles a hard limiter; however, its advantage is in its ability to deal properly with moderate outlier situations without discarding data on one hand, and without compromising performance in "benign" (no outliers) situations (as does the optimal linear estimator) on the other hand.

We demonstrated via error analysis for the case of two mixture components, that as may be expected, the attainable improvement in performance over the optimal linear estimator becomes more significant as the variance ratio of the two components increases. However, an optimum is attained as a function of the outlier probability and SNR.

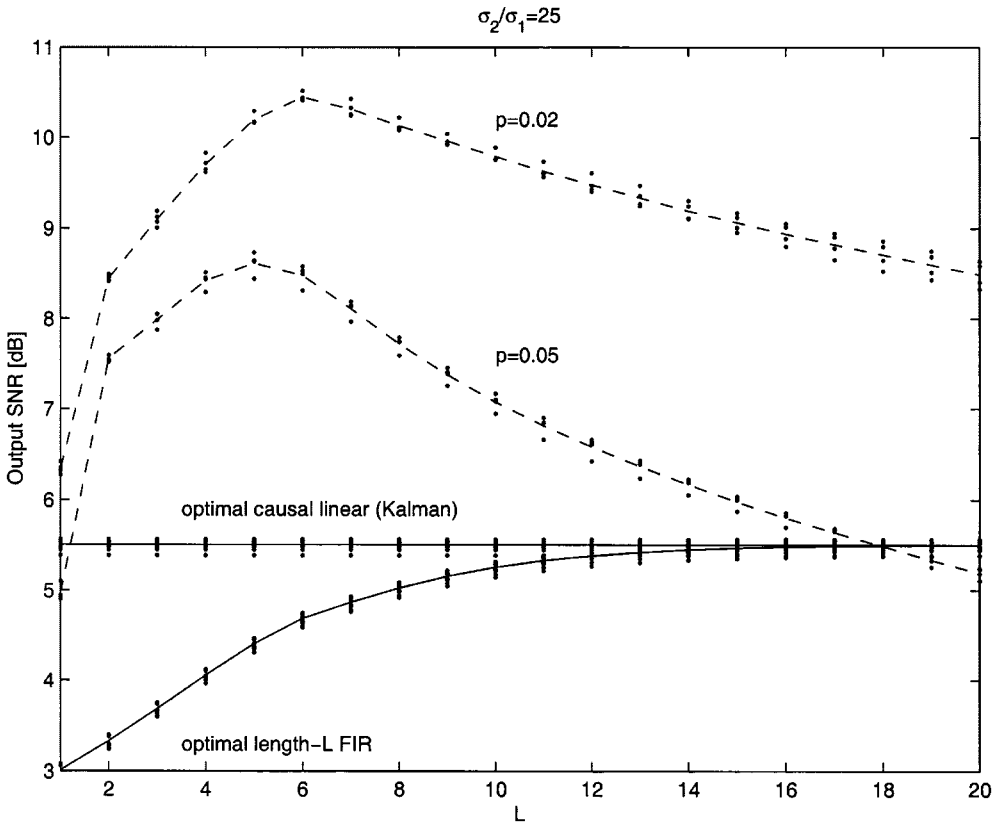

Fig. 5. SNR at the output of the optimal linear filters and proposed ZMGM filter versus the memory length L, for two outlier probabilities p. Solid lines denote the theoretical values for the optimal linear Kalman and FIR filters. Simulations results are indicated by dots, each representing a single trial over a signal of length 250 000 (four trials per test point). The same data was used by all filters for all L-s. The Kalman filter results are repeated at each L for comparison only, since L is irrelevant to the Kalman filter.

We then extended the results to the case of filtering correlated time series. We presented a suboptimal finite-memory filter for estimating a stationary Gaussian signal from measurements corrupted by colored ZMGM noise. Under the assumption that independent occurrences of outliers are usually sufficiently far apart, the filter is nearly optimal. Using simulation results, its superiority with respect to the optimal linear filter, as well as its sensitivity to the memory length L, were demonstrated.

In general, Gaussian mixture models can be used in applications other than signal denoising, involving estimation with impulsive signals, such as channel estimation and equalization, source separation, bearing estimation, time-delay estimation, etc. Often in such applications, the ZMGM modeling tool can be used either to properly combat the undesired effects of impulsive noise, or to exploit useful impulsive properties of the target signal.

The nonlinear functions employed by using the ZMGM model are reminiscent of those generated by other nonlinear filtering methods, such as piecewise linear filters [8], piecewise polynomials [9], and threshold decomposition [10]. However, in contrast to these other methods, the proposed ZMGM filter is (nearly) optimal when the statistical model is indeed ZMGM. The other nonlinear methods are not specifically related to (nor are claimed optimal in) the context of a specific statistical model. Thus, these methods would be more robust with respect to the model assumption, but may be far from optimal when the true model is indeed a Gaussian mixture.

APPENDIX
ESTIMATION ERROR ANALYSIS

In this Appendix, we derive expressions for computing the MSE attained by the optimal estimator $\hat{x}(y) = E[x \mid y]$ in the iid case, for noise consisting of two mixture components ($M = 2$). Throughout, we shall use the notations defined in Section II for the distributions' parameters and related constants.

The implicit expression for the MSE is given by

$$\begin{aligned} E[(\hat{x}(y) - x)^2] &= E[E[(\hat{x}(y) - x)^2 \mid y]] \\ &= \sigma_x^2 - E[\hat{x}^2(y)]. \end{aligned} \tag{30}$$

Thus, evaluating the MSE involves the evaluation of $E[\hat{x}^2(y)]$, which in the case $M = 2$ reduces to

$$E[\hat{x}^2(y)] = \int_{-\infty}^{\infty} \left(\frac{\frac{p_1}{s_1}\exp\left(-\frac{y^2}{2s_1^2}\right)h_1 + \frac{p_2}{s_2}\exp\left(-\frac{y^2}{2s_2^2}\right)h_2}{\frac{p_1}{s_1}\exp\left(-\frac{y^2}{2s_1^2}\right) + \frac{p_2}{s_2}\exp\left(-\frac{y^2}{2s_2^2}\right)} \right)^2 \times y^2 f_y(y)\, dy. \tag{31}$$

To simplify the exposition, let us define the constants $\gamma_i = p_i/s_i (i = 1, 2)$ and $(1/s^2) = (1/s_1^2) - (1/s_2^2)$. Noting that the denominator in parenthesis in (31) equals $\sqrt{2\pi} f_y(y)$ and defining $g_i(y) = \gamma_i h_i \exp(-(y^2/2s_i^2))$ $(i = 1, 2)$,

we may break (31) down into three terms denoted $(1/\sqrt{2\pi})(T_{11}+2T_{12}+T_{22})$, where

$$T_{ij}=\int_{-\infty}^{\infty}\frac{g_i(y)g_j(y)}{\sqrt{2\pi}f_y(y)}y^2\,dy\triangleq t_{ij}T'_{ij},\qquad i,j=1,2. \quad (32)$$

Here, $t_{ij}=\gamma_i h_i\gamma_j h_j/\gamma_2$ and $T'_{ij}=\int_{-\infty}^{\infty}(\exp(-a_{ij}y^2)/(1+b\exp(-cy^2))y^2\,dy$ where $b=\gamma_1/\gamma_2, c=(1/2s^2)$ and $a_{ij}=(1/2)((1/s_i^2)+(1/s_j^2)-(1/s_2^2))$ for $i,j=1,2$.

Evaluating T_{ij} amounts to evaluating T'_{ij}, for which there is no known closed-form solution. Nevertheless, we may get rid of the denominator in T'_{ij} by exploiting the relations

$$\frac{1}{1+z}=\begin{cases}\sum_{n=0}^{\infty}(-1)^n z^n, & |z|<1\\ \sum_{n=1}^{\infty}(-1)^{n+1}z^{-n}, & |z|>1\end{cases} \quad (33)$$

as follows: define $z=b\exp(-cy^2)$. Assuming $p_1>p_2$ (which is a common assumption in an outlier situation), we have $b>1, c>0$, and we may, therefore, use $y_{\rm th}\triangleq\sqrt{(1/c)\log(b)}$ to partition the integration domain into two regions: D_1:$\{y$:$|y|>y_{\rm th}\}$ and D_2:$\{y$:$|y|<y_{\rm th}\}$ in which $|z|<1$ and $|z|>1$, respectively. We may now evaluate T'_{ij} using

$$\int_{D_k}y^2\exp(-\alpha y^2)\,dy=\frac{1}{\alpha\sqrt{\alpha}}\cdot\begin{cases}\frac{\sqrt{\pi}}{2}\mathrm{erfc}(\sqrt{\alpha}y_{\rm th})+\sqrt{\alpha}y_{\rm th}\exp(-\alpha y_{\rm th}^2), & k=1\\ \frac{\sqrt{\pi}}{2}\mathrm{erf}(\sqrt{\alpha}y_{\rm th})-\sqrt{\alpha}y_{\rm th}\exp(-\alpha y_{\rm th}^2), & k=2\end{cases} \quad (34)$$

where $\mathrm{erf}(t)=(2/\sqrt{\pi})\int_0^t\exp(-w^2)\,dw$ and $\mathrm{erfc}(t)=1-\mathrm{erf}(t)$.

Let us also define $Q_k(\alpha)\ k=1,2$ as the two possible results of (34). Combining (33) and (34), we may express T'_{ij} as an infinite sum as follows:

$$T'_{ij}=\sum_{n=0}^{\infty}(-1)^n\left(\frac{\gamma_1}{\gamma_2}\right)^n Q_1(a_{ij}+nc)-\sum_{n=1}^{\infty}(-1)^n\left(\frac{\gamma_1}{\gamma_2}\right)^{-n}Q_2(a_{ij}-nc). \quad (35)$$

The expression $a_{ij}-nc$ in the second term of (35) becomes negative for values of n beyond a certain threshold. When $Q_2(\cdot)$ obtains a negative value as its argument, $\mathrm{erf}(\cdot)$ of an imaginary argument has to be computed. The definition for $\mathrm{erf}(t)$ stated below (34) holds true for any complex argument t. However, standard tables and routines for evaluating $\mathrm{erf}(t)$ are usually available for real valued t-s only (e.g., in MATLAB). We, therefore, provide below the following sum for evaluating $\mathrm{erf}(\tilde{t})$ for imaginary values of $\tilde{t}=it$ (here $i=\sqrt{-1}$ and t is real-valued):

$$\mathrm{erf}(it)=\frac{i}{\pi}\left[t+2\sum_{n=1}^{\infty}\frac{1}{n}\exp\left(-\frac{1}{4}n^2\right)\sinh(nt)\right] \quad (36)$$

(see, e.g., [7, p. 299]).

Thus, the MSE of $\hat{x}(y)$ can be calculated analytically to within arbitrary precision by taking sufficiently many terms in (35).

REFERENCES

[1] J. Bodenschatz and C. Nikias, "Symmetric alpha-stable filter theory," *IEEE Trans. Signal Processing*, vol. 45, pp. 2301–2306, Sept. 1997.

[2] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[3] T. Kasparis and J. Lane, "Suppression of impulsive disturbances from audio signals," *Proc. IEEE*, vol. 73, pp. 433–481, Mar. 1985.

[4] C. J. Masreliez, "Approximate non-Gaussian filtering with linear state and observation relations," *IEEE Trans. Automat. Contr.*, vol. 20, pp. 107–110, Feb. 1975.

[5] M. Shao and C. Nikias, "Signal processing with fractional lower order moments: Stable processes and their applications," *Proc. IEEE*, vol. 81, pp. 986–1009, July 1993.

[6] H. W. Sorenson and D. L. Alspach, "Recursive bayesian estimation using Gaussian sums," *Automatica*, vol. 7, pp. 465–479, 1971.

[7] I. Stegun and M. Abramowitz, *Handbook of Mathematical Functions*, NY: Dover, 1972.

[8] E. A. Heredia and G. R. Arce, "Piecewise linear system modeling based on a continuous threshold decomposition," *IEEE Trans. Signal Processing*, vol. 44, pp. 1440–1453, June 1996.

[9] ——, "Nonlinear filters based on combinations of piecewise polynomials with compact support," *IEEE Trans. Signal Processing*, vol. 48, pp. 2850–2863, Oct. 2000.

[10] G. R. Arce, "Microstatistics in signal decomposition and the optimal filtering problem," *IEEE Trans. Signal Processing*, vol. 40, pp. 2669–2682, Nov. 1992.

[11] S. M. Verbout, J. M. Ooi, J. T. Ludwig, and A. V. Oppenheim, "Parameter estimation for autoregressive Gaussian-mixture processes: The EMAX algorithm," *IEEE Trans. Signal Processing*, vol. 46, pp. 2744–2756, Oct. 1998.

Yonina C. Eldar (S'98) received the B.Sc. degree in physics, and the B.Sc. degree in electrical engineering from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1995 and 1996, respectively. She is currently working toward the Ph.D. degree in electrical engineering at Massachusetts Institute of Technology (MIT), Cambridge.

From 1992 to 1996, she was in the program for outstanding students at TAU. Since 1998, she has been with the Digital Signal Processing Group at MIT. She has served as a Teaching Assistant for classes in linear systems, digital signal processing, parameters estimation, and statistical signal processing. Her current research interests are in the general areas of signal processing and quantum detection.

Ms. Eldar was awarded the Rosenblith Fellowship for study in electrical engineering at MIT in 1998. She is currently the recipient of an IBM Research Fellowship.

Arie Yeredor (M'99) was born in Haifa, Israel, in 1963. He received the B.Sc. degree in electrical engineering (*summa cum laude*) and the Ph.D. degree from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1984 and 1997, respectively.

From 1984 to 1990, he was with the Israeli Defense Forces (Intelligence Corps), in charge of advanced research and development activities in the fields of statistical and array signal processing. Since 1990, he has been with NICE Systems Inc., Ra'anana, Israel, where he holds a consulting position in the fields of speech and audio processing, video processing, and emitter location algorithms. He is currently a Faculty member in the Department of Electrical Engineering—Systems at TAU, where he teaches courses in parameters estimation and statistical and digital signal processing. His research interests include estimation theory, statistical signal processing, and blind source separation.

Dr. Yeredor has been awarded the Best Lecturer of the Faculty of Engineering Award for three consecutive years at TAU.