different. The average run time (for the minimization of the objective function) was reduced to 4.2 s using the MUSIC initial estimates. The numerical experiment was performed under the following conditions:

1) The array is linear with an intersensor spacing of $\frac{1}{2}$ in electrical units (i.e., wavelengths).
2) The array consists of eight sensors. Two targets of equal strength are placed at ± 45 from broadside.
3) The noise $v(n)$ is a zero mean complex Gaussian such that $E[v(n)v^{+}(m)] = \delta_{mn}\sigma_n^2 I$.
4) It is clear that for the signal vectors $a(n)$, the following holds: $E[a(n)a^{+}(m)] = \delta_{mn}\sigma_s^2 I$.

The signal-to-noise ratio is thus given by $20\log(\sigma_s/\sigma_n)$. For a given signal-to-noise ratio, N snapshots of the received data are taken. This procedure is repeated 100 times, and the mean square error in the estimate is calculated. The mean square error shown is the average of the mean square errors for the different angles of arrival. The standard deviation shown is the maximum standard deviation recorded for any of the estimates. The results are shown in Figs. 1–4.

From Fig. 1, it is clear that the standard deviation of the estimate drops to acceptably low levels ($\leq$ 0.03 rads) if more than 10 snapshots are used. It is also clear that increasing the number of samples (P) used to estimate the noise covariance matrix leads to improved standard deviations for any number of snapshots (N).

From Fig. 2, it is clear that the mean square error drops sharply as N is increased to ten snapshots. Again, increasing P leads to improved mean square errors. In Fig. 3, the mean square error is shown for decreasing signal-to-noise ratios. The standard deviation versus decreasing signal-to-noise ratios is plotted in Fig. 4. These figures show that the algorithm degrades gently in the presence of noise. This should be seen as a direct consequence of the structure of the algorithm, i.e., of the fact that an experimental estimate of the noise is used to compute the estimate.

From the simulations, it is clear that the optimal N is at approximately 20 snapshots. If more than 20 snapshots are used, only minor reductions in standard deviation and mean square error are achieved. This does not justify the extra computational time that would be incurred.

IV. Conclusions

In this correspondence, a Bayesian approach to the estimation problem in the presence of arbitrary noise has been adopted. The resulting algorithm is optimal in the *maximum a posteriori* sense. The algorithm differs from recent approaches to the problem, which also employ a Bayesian approach [1]. These methods assume that the noise is completely unknown. In the present case, an experimental estimate of the noise is used. The resulting objective function uses the estimate, together with the number of samples used to determine the estimate as input parameters. Thus, the previously derived estimator [1] reduces to this estimation rule for the case that no experimental estimate is available.

The algorithm has been studied using numerical simulations. The results show that the resulting estimator is robust with respect to noise and degrades gently for increasing signal-to-noise ratios. Increasing the number of samples used to determine the experimental estimate of the noise covariance matrix decreases both the mean square error and the standard deviation of the estimate. It is for this reason that this estimator should be considered to be an improvement on existing techniques.

References

[1] K. M. Wong, J. P. Reilly, Q. Wu, and S. Qiao, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 40, p. 2007, 1992; *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 40, p. 2018, 1992.
[2] A. J. Willis and R. De Mello Koch, *Electron. Lett.*, vol. 28, p. 358, 1992.
[3] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1986.
[4] S. S. Reddi, *IEEE Trans. Aerosp. Electon. Syst.*, vol. AES-15, p. 1, 1979.
[5] R. O. Schmidt, *IEEE Trans. Antennas Propagat.*, vol. AP-34, p. 276, 1986.
[6] M. I. Skolink, *Radar Handbook.* New York: McGraw-Hill, 1970, pp. 26-1–26-9.
[7] H. L. Van Trees, *Detection, Estimation and Modulation Theory Part I.* New York: Wiley, 1968.
[8] A. J. Willis, B. Spear, A. Klopper, and R. De Mello Koch, in *Proc. IEEE APS/URSI Joint Symp.*, Ann Arbor MI, June 28–July 2, 1993, vol. 3, pp. 1876–1878.
[9] S. U. Pillai, *Array Signal Processing*. New York: Springer-Verlag, 1989.
[10] N. R. Goodman, *Ann. Math. Stat.*, vol. 34, p. 152, 1963.
[11] M. D. Srinath and P. K. Rajasekaran, *An Introduction to Statistical Signal Processing with Problems*. New York: Wiley, 1979.
[12] R. Maartens, unpublished notes.
[13] R. De Mello Koch, unpublished notes.
[14] I. Ziskind and M. Wax, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, p. 1553, 1988.

Robust Direction-of-Arrival Estimation in Non-Gaussian Noise

Yasemin Yardımcı, A. Enis Çetin, and James A. Cadzow

***Abstract*—In this correspondence, a nonlinearly weighted least-squares method is developed for robust modeling of sensor array data. Weighting functions for various observation noise scenarios are determined using maximum likelihood estimation theory. Computational complexity of the new method is comparable with the standard least-squares estimation procedures. Simulation examples of direction-of-arrival estimation are presented.**

***Index Terms*—Antenna arrays, direction-of-arrival estimation, impulsive noise, maximum likelihood estimation.**

I. Introduction

In many signal modeling problems, including array signal processing, the key issue is to estimate the parameters of some basis signals from the observations. Array processing techniques based on minimization of the squared model error (ℓ_2-norm) have received considerable attention [1]–[3]. The popularity of least squares (LS) based methods is justified by the fact that their solutions are equivalent to that of maximum likelihood (ML) for signals embedded in independent identically distributed (i.i.d.) Gaussian noise.

Manuscript received July 11, 1996; revised May 8, 1997. The associate editor coordinating the review of this paper and approving it for publication was Prof. Hagit Messer-Yaron.

Y. Yardımcı and A. E. Çetin are with the Department of Electrical Engineering, Bilkent University, Ankara, Turkey.

J. A. Cadzow is with the Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN 37325 USA.

Publisher Item Identifier S 1053-587X(98)03277-2.

Unfortunately, the performance of ℓ_2-norm based methods deteriorate in nonstationary or non-Gaussian noise environments, especially when the data contains outliers [4]–[9]. Such deviations may result from various reasons including changing conditions through the course of operation and the presence of impulsive noise. Examples of application areas with non-Gaussian environments are sonar [10], [11], radar [12], and communication systems [13]. Estimation schemes based on norms other than two (ℓ_p-norm, $1 \leq p < 2$) are suggested as robust alternatives; however, they are not as efficient as the standard LS in Gaussian noise [4]. Furthermore, their solutions are nonlinear functions of the data; therefore, increase the computational complexity of an optimization solution technique. In this paper, a direction-of-arrival (DOA) estimation scheme, which is robust and yet has the computational advantages of the standard least squares method, is described. Robustness with respect to modeling errors in the noise distribution is achieved by introducing a nonlinearity that weights the squared error terms corresponding to snapshots obtained at different time instants. In our modeling, the noise observations are assumed to be i.i.d. across the antennas for a given time instant. Nonlinear weighting functions for various observation noise scenarios are determined by ML estimation.

II. Signal Model and Robust Parameter Estimation

In a passive sensor array, the signal generated on the p sensors can be described by a $p \times 1$ vector

$$\underline{x}(t) = \sum_{i=1}^{m} a_i(t)\underline{s}_i(\underline{\theta}_i) + \underline{w}(t) \tag{1}$$

where

- $\underline{s}_i(\underline{\theta}_i)$ steering vectors that are dependent on the parameter vectors $\underline{\theta}_i$;
- $a_i(t)$ amplitude of the ith signal vector;
- $\underline{w}(t)$ measurement noise vector;
- m number of sources.

The parameter vector $\underline{\theta}_i$ may be composed of the azimuth and the elevation angles of the ith impinging plane wave for sources in the far field or coordinates of the ith point source in three-dimensional space for sources in the near field. The sensor output vectors in (1) can be rewritten in the compact form as

$$\underline{x}(t) = S(\underline{\theta})\underline{a}(t) + \underline{w}(t) \tag{2}$$

where $\underline{\theta} = [\underline{\theta}_1^T\ \underline{\theta}_2^T\ \cdots\ \underline{\theta}_m^T]^T$ is the $q \times 1$ composite unknown parameter vector that includes the DOA's, $\underline{a}(t) = [a_1(t)\, a_2(t)\, \cdots\, a_m(t)]^T$ is the amplitude vector, and $S(\underline{\theta})$ is the steering matrix whose columns are formed by the m steering vectors $\underline{s}_i(\underline{\theta}_i)$. The parameter estimation problem is one of estimating the vector $\underline{\theta}$ and the unknown amplitudes $\underline{a}(t)$ from the observed signal $\underline{x}(t)$. In practice, the observed signal $\underline{x}(t)$ is sampled so that we have only a sequence of N vectors $\{\underline{x}(t_1), \underline{x}(t_2), \cdots, \underline{x}(t_N)\}$ from which $\underline{\theta}$ and $\underline{a}(t_n)$ should be estimated. Robust ML-based DOA estimation problem has also been considered in [14] and [15]. The former models the amplitude vectors as samples from a random process, thereby adhering to the unconditional or stochastic model of array snapshot vectors. The latter reference, which employs the conditional or deterministic model as we do in this study, assumes the amplitude vectors $\underline{a}(t_n)$ are known *a priori*.

The standard least squares error (LSE) method is based on minimization of the squared error criterion

$$e[\underline{a}(t_1), \underline{a}(t_2), \cdots, \underline{a}(t_N), \underline{\theta}] = \sum_{n=1}^{N} ||\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)||^2 \tag{3}$$

with respect to $\underline{\theta}$ and $\underline{a}(t_n)$ for $n = 1, 2, \cdots, N$. In a scenario where there are m sources and N time samples, estimating the N $m \times 1$ complex amplitude vectors $\underline{a}(t_n)$ and the $q \times 1$ composite unknown parameter vector $\underline{\theta}$ requires a multidimensional search in $2Nm + q$-dimensional space. Furthermore, the vector $\underline{\theta}$ enters the criterion (3) in a nonlinear manner so that a closed-form solution for $\underline{a}(t_n)$ and $\underline{\theta}$ is not feasible. However, for a given $\underline{\theta}$, the optimum $\underline{a}^o(t_n)$ is given by

$$\underline{a}^o(t_n) = [S^*(\underline{\theta})S(\underline{\theta})]^{-1}S^*(\underline{\theta})\underline{x}(t_n) = S^\dagger(\underline{\theta})\underline{x}(t_n) \tag{4}$$

where $S^*(\underline{\theta})$ and $S^\dagger(\underline{\theta})$ are the complex conjugate transpose and the pseudo-inverse of the matrix $S(\underline{\theta})$, which is tacitly assumed to have a full column rank. By substituting the optimum amplitudes $\underline{a}^o(t_n)$ given by (4) for $t = t_1, t_2, \cdots, t_N$ in (3), a modified criterion

$$e[\underline{a}^o(t_1), \underline{a}^o(t_2), \cdots, \underline{a}^o(t_N), \underline{\theta}] = \sum_{n=1}^{N} ||[I - P(\underline{\theta})]\underline{x}(t_n)||^2 \tag{5}$$

is obtained. In this expression, $P(\underline{\theta}) = S(\underline{\theta})S^\dagger(\underline{\theta})$ is the projection matrix onto the column space of $S(\underline{\theta})$. The minimization problem (5) requires a search for $\underline{\theta}$ in only q-dimensional space. Thus, the two-step procedure significantly reduces the computational complexity of the problem. Moreover, Golub and Pereyra [16] showed that the global minimizer $\underline{\theta}^o$ of (5) is also the global minimizer of (3).

In order to achieve a robust estimate, we employ a generalization of the standard LS as specified by

$$e[\underline{a}(t_1), \underline{a}(t_2), \cdots, \underline{a}(t_N), \underline{\theta}] = \sum_{n=1}^{N} \psi[||\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)||^2] \tag{6}$$

where $\psi(.)$ is a nonlinear weighting function defined on the positive real axis with argument as the squared error $r(t_n) = \|\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)\|^2$. The standard LS corresponds to the choice of $\psi(.)$ as the identity function $\psi(r) = r$. The robust estimates of $\underline{\theta}$ and $\underline{a}(t)$ are obtained by minimizing (6). The gradient of (6) with respect to the unknown amplitude vectors $\underline{a}(t_n)$ is given by

$$\begin{aligned}\nabla_{\underline{a}(t_n)} e[&\underline{a}(t_1), \underline{a}(t_2), \cdots, \underline{a}(t_N), \underline{\theta}] \\ &= 2\dot{\psi}[\|\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)\|^2] \\ &\quad \cdot \{[S^*(\underline{\theta})S(\underline{\theta})]\underline{a}(t_n) - S^*(\underline{\theta})\underline{x}(t_n)\} \quad \text{for } n = 1, 2, \cdots, N\end{aligned} \tag{7}$$

where $\dot{\psi}$ denotes the derivative of ψ with respect to its argument. For monotonically increasing weighting functions ψ, $\dot{\psi} > 0$, and the optimum amplitudes obtained through the minimization of the nonlinear squared error criterion in (6) are given by $\underline{a}^o(t_n) = S^\dagger(\underline{\theta})\underline{x}(t_n)$ for $n = 1, 2, \cdots, N$, which is identical to the solution of the standard LS problem described by (3). Another desirable property of the nonlinear weighting function ψ is that it increases more slowly than $\psi(r) = r$ for large values of r so that the outliers are deemphasized. Substituting these optimum amplitudes in (6) yields

$$e[\underline{a}^o(t_1), \underline{a}^o(t_2), \cdots, \underline{a}^o(t_N), \underline{\theta}] = \sum_{n=1}^{N} \psi\{||[I - P(\underline{\theta})]\underline{x}(t_n)||^2\} \tag{8}$$

whose solution again requires only a q-dimensional search. The minimization of (8) with respect to $\underline{\theta}$ can be performed through a nonlinear programming algorithm.

Conventionally, noise vectors $\underline{w}(t_n)$, $n = 1, 2, \cdots, N$ are assumed to be complex valued, zero-mean Gaussian vectors with the covariance matrix $\sigma^2 I$, where σ^2 is an unknown scalar [2], [3]. In this case, the deterministic ML estimator turns out to be equivalent to the standard LS solution of (3) with optimum weighting function $\psi(r) = r$. In the following subsections, nonlinear weighting functions

for various observation noise scenarios are obtained by ML estimation theory.

A. ML Estimation in the Presence of Gaussian Noise with Changing Variance

The variance of the Gaussian observation noise may change in time. In that case, $\underline{w}(t_n) \sim \mathcal{N}(0, \sigma_n^2 I)$, where $\sigma_n^2 I$ is the noise covariance matrix at time t_n for $n = 1, 2, \cdots, N$. Then, the LS estimator obtained by minimizing (3) is no longer optimal in the ML sense, and it is susceptible to severe degradation. If the noise samples are temporally independent, the joint density function of the observed data is given by

$$f[\underline{x}(t_1), \underline{x}(t_2), \cdots, \underline{x}(t_N)] = \prod_{n=1}^{N} \frac{1}{\pi \det[\sigma_n^2 I]} \exp\left[-\frac{1}{\sigma_n^2}||\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)||^2\right] \quad (9)$$

where $\det[\sigma_n^2 I] = \sigma_n^{2p}$. The log-likelihood function is

$$\mathcal{L} = -\sum_{n=1}^{N}\left[\log\ \pi + p\ \log(\sigma_n^2) + \frac{1}{\sigma_n^2}||\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)||^2\right]. \quad (10)$$

In order to obtain the ML estimates of $\underline{\theta}$ and $\underline{a}(t)$, this log-likelihood function is maximized with respect to unknown parameters $\underline{\theta}$, $\underline{a}(t_n)$, and σ_n^2, $n = 1, 2, \cdots, N$. Maximization of (10) can be achieved by minimizing an expression of the form (6) with respect to unknown parameters $\underline{\theta}$, $\underline{a}(t_n)$, and σ_n^2 for $n = 1, 2, \cdots, N$

$$\mathcal{L} = \sum_{n=1}^{N} \psi[||\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)||^2] \quad (11)$$

where

$$\psi[r(t_n)] = \log\ \pi + p\ \log(\sigma_n^2) + \frac{r(t_n)}{\sigma_n^2}. \quad (12)$$

For fixed $\underline{\theta}$ and $\underline{a}(t_n)$, the ML estimate of σ_n^2 is given by

$$\hat{\sigma}_n^2 = \frac{1}{p}\|\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)\|^2 = \frac{r(t_n)}{p} \quad \text{for } n = 1, 2, \cdots, N. \quad (13)$$

Substituting (13) back into (12), the nonlinearity

$$\psi[r(t_n)] = \log\ \pi + p\ \log\ \frac{r(t_n)}{p} + p \quad (14)$$

is obtained. This corresponds to the minimization problem

$$\min_{\underline{\theta},\ \underline{a}(t_n)} \sum_{n=1}^{N} \log\ ||\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)||^2 \quad (15)$$

when the constant terms in (14) are ignored. The difference between (3) and (15) is the nonlinear weighting function $\psi(r) = \log(r)$. Weighting the error terms $\|\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)\|^2$ in a logarithmic fashion provides robustness with respect to changes in the noise variance. Notice that the $\psi(r) = \log(r)$ function gives less emphasis to the high-valued outlying samples and more emphasis to the small error terms compared with the $\psi(r) = r$, which is the weighting function of the standard LS estimation. Minimizing (15) with respect to the unknown amplitudes $\underline{a}(t_n)$ and substituting the optimum solutions obtained as $\underline{a}^o(t_n) = S^\dagger(\underline{\theta})\underline{x}(t_n)$ in (15) yields the q-dimensional optimization problem

$$\min_{\underline{\theta}} \sum_{n=1}^{N} \log\ ||[I - P(\underline{\theta})]\underline{a}(t_n)||^2 \quad (16)$$

which is equivalent to (8) for $\psi(r) = \log(r)$.

B. Noise with a Spherically Symmetric Distribution

In this case, we assume that the noise samples $w(t_n)$ are from a spherically symmetric distribution described by a multivariate probability density function (pdf) of the form

$$f_{\underline{w}}(\underline{w}) \alpha \frac{1}{b^{2p}} \mathrm{g}\left(\frac{\underline{w}^*\underline{w}}{b^2}\right) \quad (17)$$

so that the pdf is only a function of the Euclidean norm of the random vector $\underline{w}$. The scale parameter b controls the spread of the univariate pdf and is the standard deviation for the Gaussian distribution. The class of spherically symmetric distributions include the multivariate Gaussian distributions with covariance matrix $\sigma^2 I$[1] and the multivariate student's t distribution, which has the Cauchy distribution as a special case. When the noise samples are from a spherically symmetric distribution and are temporally independent, the joint density function is given by

$$f[\underline{x}(t_1), \underline{x}(t_2), \cdots, \underline{x}(t_N)] = \text{const.} \prod_{n=1}^{N} \frac{1}{b^{2p}} \mathrm{g}\left[\frac{1}{b^2}||\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)||^2\right]. \quad (18)$$

The corresponding negative of the log-likelihood function (constant terms ignored)

$$\mathcal{L} = -\sum_{n=1}^{N} \log\ \left\{\mathrm{g}\left[\frac{1}{b^2}||\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)||^2\right]\right\} \quad (19)$$

is to be minimized with respect to the amplitudes $\underline{a}(t_1), \underline{a}(t_2), \cdots, \underline{a}(t_N)$ and the unknown parameter vector $\underline{\theta}$. If the univariate function $\mathrm{g}(r)$ is a unimodal function with unbounded support, then the optimum amplitudes are, again, a linear function of the data[2] as $\underline{a}^o(t_n) = S^\dagger(\underline{\theta})\underline{x}(t_n)$ for $n = 1, 2, \cdots, N$. The weighting function ψ is

$$\psi(r) = -\log\ \left[\mathrm{g}\left(\frac{r}{b^2}\right)\right]. \quad (20)$$

For the multivariate Gaussian distribution, $\psi(r) = -\log[\mathrm{g}(r/b^2)] = r/b^2$, as expected. For the p-variate Cauchy distribution with pdf $f[\underline{x}(t_1), \underline{x}(t_2), \cdots, \underline{x}(t_N)] = \prod_{n=1}^{N}(c/b^{2p})[1 + (1/b^2)\|\underline{x}(t_n) - S(\underline{\theta})\underline{a}(t_n)\|^2]^{-[(1/2)(p+1)]}$, the optimum nonlinear weighting function given by (20) is obtained as $\psi(r) = \log(1 + r/b^2)$. Recently, this weighting function was also obtained in [18] within the context of α-stable distributions.

C. ML Estimation in Gaussian-Mixture Noise

Another deviation from the model assumptions of [2] and [3] may occur if noise vectors at some time instants have much higher variance than the others. A commonly employed model for such deviations from the nominal model of noise samples $\underline{w}(t_n)$ is the ϵ-contaminated Gaussian distribution whose cumulative distribution function (cdf) is given by

$$F_{\underline{w}}(\underline{w}) = (1 - \epsilon)\Phi(\underline{w};\ \sigma_1^2 I) + \epsilon\Phi(\underline{w};\ \sigma_2^2 I) \quad (21)$$

where $\Phi(\underline{w};\ I)$ is the cdf of the zero mean Gaussian vector $\underline{w}$ with the covariance matrix I, $\epsilon \in [0, 1]$, and σ_1^2 and σ_2^2 are the variances of the samples originating from the nominal and the contaminating distributions, respectively. Typically, ϵ is a number close to zero, and $\sigma_1^2 \ll \sigma_2^2$. The second term in (21) models the outlying observations. By carrying out an ML analysis as above, we get

$$\psi(r) = \log(\pi) - \log\left[\frac{1 - \epsilon}{\sigma_1^{2p}} \exp\left(-\frac{r}{\sigma_1^2}\right) + \frac{\epsilon}{\sigma_2^{2p}} \exp\left(-\frac{r}{\sigma_2^2}\right)\right] \quad (22)$$

[1] A random vector with an arbitrary covariance matrix Σ belongs to the class of elliptically symmetric distributions and can be converted to a spherically invariant distribution by an affine transformation.

[2] In fact, it is shown in [17] that for multivariate distributions with finite variance, linear regressions imply spherical symmetry.

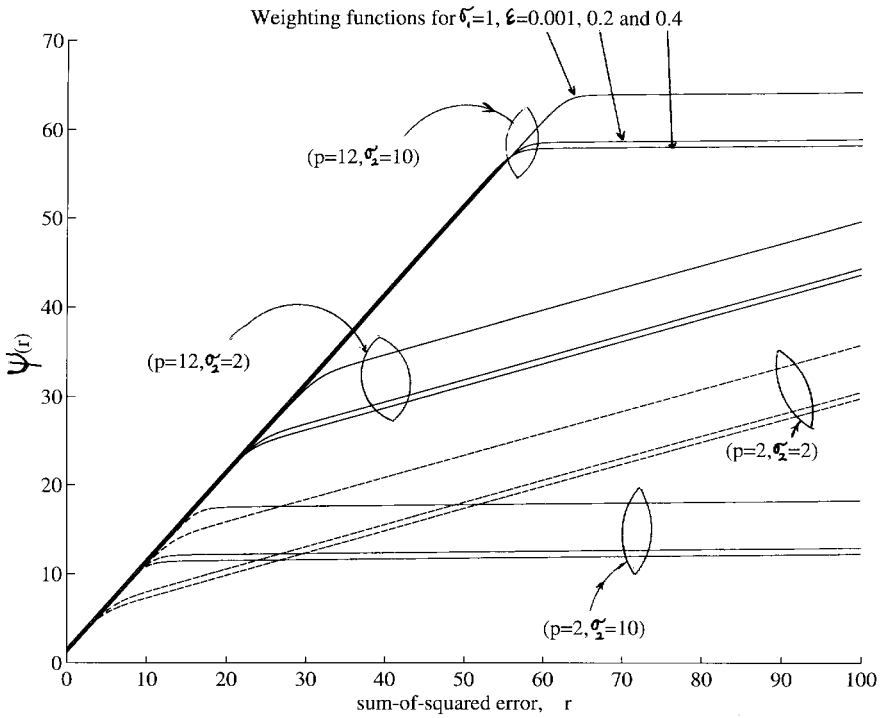

Fig. 1. Optimum ψ functions for ϵ-contaminated distributions with the nominal distribution standard deviation $\sigma_1 = 1$. The parameter p stands for the number of sensors. σ_2 is the standard deviation of the contaminating distribution, and for each selection of p and σ_2, the nonlinearity $\psi(r)$ is plotted for three different contamination rates $\epsilon = 0.001, 0.2$, and 0.4.

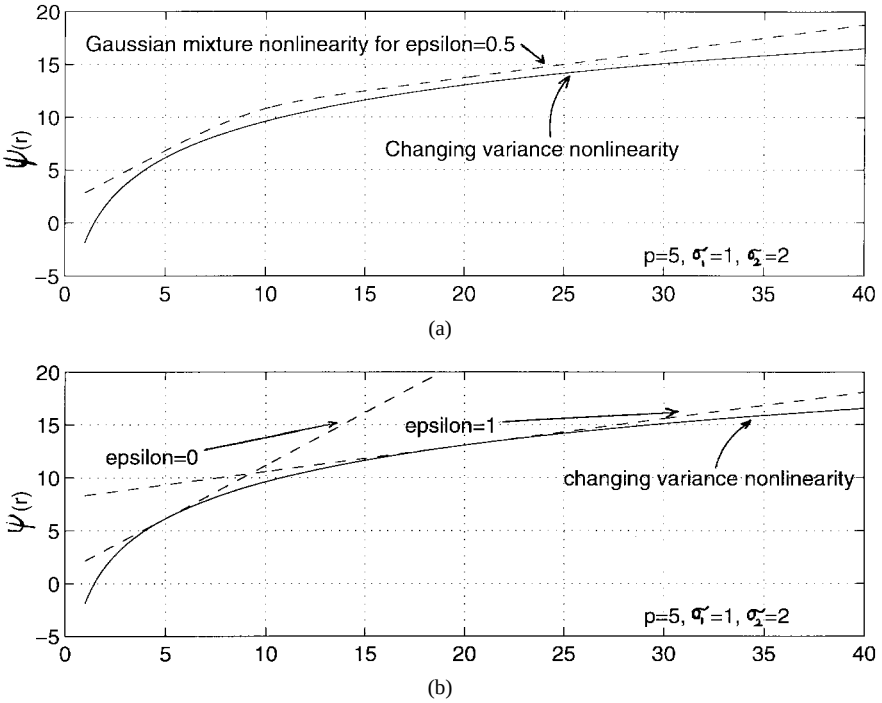

Fig. 2. Comparison of the nonlinear weighting functions for the cases of changing variance and Gaussian mixture noise with contamination rate ϵ. (a) $\epsilon = 0.5$. (b) $\epsilon = 0$ and $\epsilon = 1$.

as the weighting function. A plot of this function is shown in Fig. 1 for various number of sensors p, contamination rates ϵ, and noise variances σ_1^2 and σ_2^2.

For any given σ_1^2, σ_2^2, and r, the optimum nonlinear weighting function (22) takes values equal or greater than that of the optimum nonlinear weighting function for the "changing variance" case (14), as shown in Fig. 2(a). Moreover, the two curves are tangent to each other when $r(t_n) = p\sigma_1^2$ and $r(t_n) = p\sigma_2^2$ for $\epsilon = 0$ and $\epsilon = 1$, respectively. This is demonstrated in Fig. 2(b). It turns out that the optimum nonlinear weighting for the "changing variance" case yield estimates similar to those of weighting with (22). The nonlinear weighting function (22) can be effectively approximated

by two intersecting lines so that

$$\psi(r) = \begin{cases} \dfrac{r}{\sigma_1^2} + k_1, & \text{for } 0 \le r \le r_o \\ \dfrac{r}{\sigma_2^2} + k_2, & \text{for } r_o < r < \infty \end{cases} \tag{23}$$

where $k_1 = \log(\pi) + 2p\,\log(\sigma_1)$, $k_2 = \log(\pi) + 2p\,\log(\sigma_2) - \log\,\epsilon$, and $r_o = (k_1 - k_2)(1/\sigma_1^2 - 1/\sigma_2^2)^{-1}$. Hence, the optimum ψ function behaves like a linearly weighted LS method with weight inversely proportional to σ_1^2 for small values of r, whereas the weight is inversely proportional σ_2^2 for large values of r. The effect of the contamination rate ϵ is to shift the second line by an amount of $\log(\epsilon)$ so that the smaller the contamination, the more the weighting function ψ behaves likes the standard LS.

D. Unknown Distribution Case

If there is no prior information regarding the contaminating distribution, then a heuristically selected nonlinear weighting function may be used. The effect of outlying samples can be reduced by selecting ψ as a function that saturates for increasing positive values of its argument. Among a multitude of such functions, we used the sigmoid function of neural networks $\psi(r) = a[1 - \exp(-\beta r)]$ and $\psi(r) = a[1 - \exp(-\beta r)]/[1 + \exp(-\beta r)]$ as well as a soft-limiter similar to Huber's estimator [4]

$$\psi(r) = \begin{cases} \dfrac{r}{\sigma_1^2} & \text{for } 0 \le r \le a\sigma_1^2 \\ a & \text{for } a\sigma_1^2 < r < \infty. \end{cases} \tag{24}$$

All of these nonlinearities exhibit linear behavior around the origin so that the samples with smaller squared-error values contribute to the criterion (6), as they would for (3). Since criterion (3) is equivalent to the ML estimation for the temporally i.i.d. Gaussian distribution, linear behavior around the origin is especially appropriate if the nominal distribution is the Gaussian. The upper bound of ψ provides robustness to the undesired outlying samples by limiting their effect to the overall cost term (6). We observed that these nonlinearities yield similar results for similar values of the upper bound a. When the noise distribution is known, the upper bound can be selected by referring to the optimum nonlinear function. If this information is not available, then the upper bound can be selected as twice the median of the sum-of-squared errors so that the effect of samples with high residual errors is limited.

E. Nonlinear Programming Method

We minimize the criterion (8) using the Gauss–Newton nonlinear programming technique. It is based on iteratively modifying the parameter to be estimated $\underline{\theta}$ by a perturbation vector $\underline{\delta}$. In order to efficiently achieve the desired minimum, a step-size scalar α is usually incorporated to adjust the perturbation vector. At the kth step of the Gauss–Newton method, the parameter vector $\underline{\theta}^k$ is updated as

$$\underline{\theta}^{k+1} = \underline{\theta}^k + \alpha_k \underline{\delta}^k. \tag{25}$$

The perturbation vector $\underline{\delta}_k$ can be shown [19] to be given by

$$\underline{\delta}^k = -\left(\sum_{n=1}^{N} \dot{\psi}\{\|[I - P(\underline{\theta}^k)]\underline{x}(t_n)\|^2\}\text{real}\{J_n^*(\underline{\theta}^k)J_n(\underline{\theta}^k)\}\right)^{-1} \cdot \sum_{n=1}^{N} \dot{\psi}\{\|[I - P(\underline{\theta}^k)]\underline{x}(t_n)\|^2\}\text{real}\{J_n^*(\underline{\theta}^k)[I - P(\underline{\theta}^k)]\underline{x}(t_n)\} \tag{26}$$

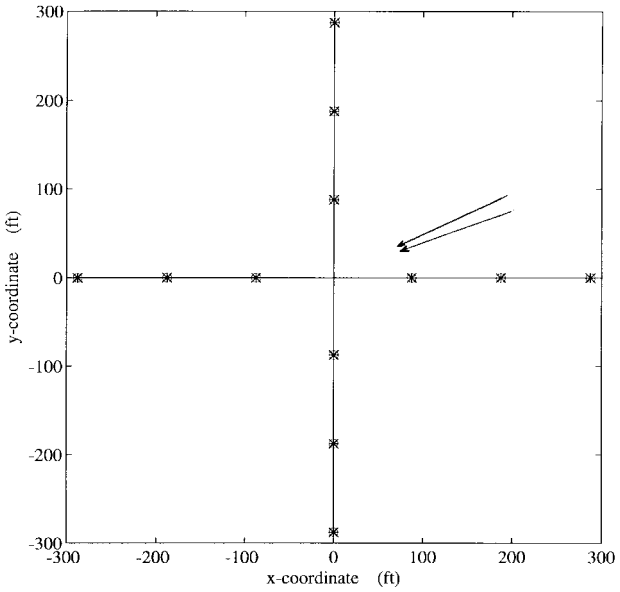

Fig. 3. Mills–Cross array with two incident signals from 21 and 25°.

TABLE I
EFFECT OF DIFFERENT WEIGHTING SCHEMES ON THE ROOT MEAN SQUARE ERRORS (DEGREES) FOR THE DOA'S UNDER CAUCHY NOISE (DOA = 25°)

$-20\log b$	LSE $\psi(r) = r$	ML $\log(\frac{r}{b^2} + 1)$	$\log(r)$	$\frac{r}{r^2+\beta_o}$
15	47.0001	0.2413	0.2418	0.3255
20	48.0533	0.1254	0.1257	0.1653
25	33.1822	0.0725	0.0726	0.0953
30	33.0244	0.0385	0.0385	0.0585
35	5.7376	0.0208	0.0208	0.0412
40	1.7866	0.0132	0.0133	0.0267
45	2.8684	0.0071	0.0071	0.0205
50	0.3258	0.0044	0.0044	0.0157

where the Jacobian matrices $J_n(\underline{\theta}^k)$ are defined as

$$J_n(\underline{\theta}^k) = \left\{ \frac{\partial}{\partial\theta_1^k}[I - P(\underline{\theta}^k)]\underline{x}(t_n) \vdots \frac{\partial}{\partial\theta_2^k}[I - P(\underline{\theta}^k)]\underline{x}(t_n) \right. \left. \vdots \cdots \vdots \frac{\partial}{\partial\theta_q^k}(I - P(\underline{\theta}^k)]\underline{x}(t_n) \right\} \quad \text{for } 1 \le n \le N \tag{27}$$

and closed-form expressions for the partial derivatives are given in [1], [16], and [19] as

$$\frac{\partial}{\partial\theta_l^k}[I - P(\underline{\theta}^k)]\underline{x}(t_n) = -\left[(I - P_{\underline{\theta}})\frac{\partial S(\underline{\theta}^k)}{\partial\theta_k^l}S(\underline{\theta}^k)^\dagger\right]\underline{x}(t_n) - \left[(I - P_{\underline{\theta}})\frac{\partial S(\underline{\theta}^k)}{\partial\theta_k^l}S(\underline{\theta}^k)^\dagger\right]^*\underline{x}(t_n). \tag{28}$$

The step-size scalar α_k is usually selected large at early iterations and reduced at later stages of the optimization procedure. A simple procedure that has been observed to be successful is to select a geometrically decreasing sequence of step sizes, i.e.,

$$\alpha_k = 1,\ \tfrac{1}{2},\ \tfrac{1}{4},\ \tfrac{1}{8},\ \cdots,\ \left(\tfrac{1}{2}\right)^i,\ \cdots \tag{29}$$

until an improving value for the updated parameter vector $\underline{\theta}^{k+1}$ is obtained.

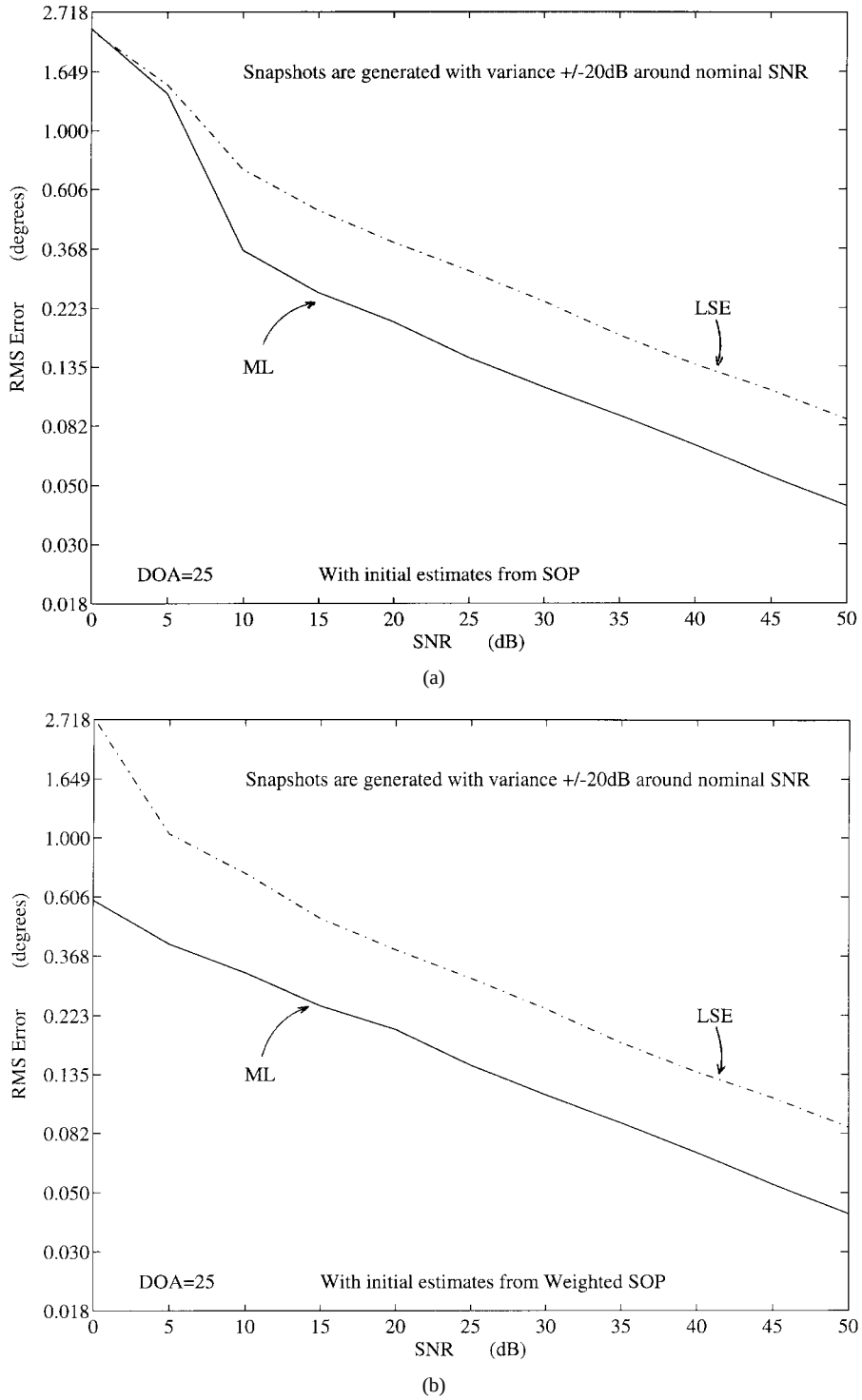

Fig. 4. Statistical comparison of the standard LS algorithm with logarithmic weighting under changing variance of ± 20 dB around nominal SNR. (a) Initial estimates are from sequential orthogonal projection algorithm. (b) Initial estimates are from weighted sequential orthogonal projection algorithm.

When the noise distribution is known, the Gauss–Newton method can be used directly as described above. For the unknown distribution case, the upper bound of the nonlinear weighting function should be determined from the data as well. In the initial iterations of the Gauss–Newton method, the residual error vectors $[I - P(\underline{\theta}^k)]\underline{x}(t_n)$ are generally large because the current estimates of the parameters are far from their actual values. In this case, a low value for the upper bound adversely affects the speed of convergence of the nonlinear programming technique. Around the vicinity of the correct value of the parameter, however, the residual errors are typically small, and hence, a predetermined upper bound may be too high to limit the effect of the outliers. We selected the upper bound as twice the median of the sum-of-squared residual errors at each step so that it adapts to the changes in the residuals throughout the optimization algorithm.

Finally, we implemented the Gauss–Newton method with a Gram–Schmidt orthogonalization step by decomposing the steering matrix $S(\underline{\theta})$ as the product of an orthonormal matrix $Q(\underline{\theta})$ and an upper matrix $R(\underline{\theta})$ as described in [1] and [19].

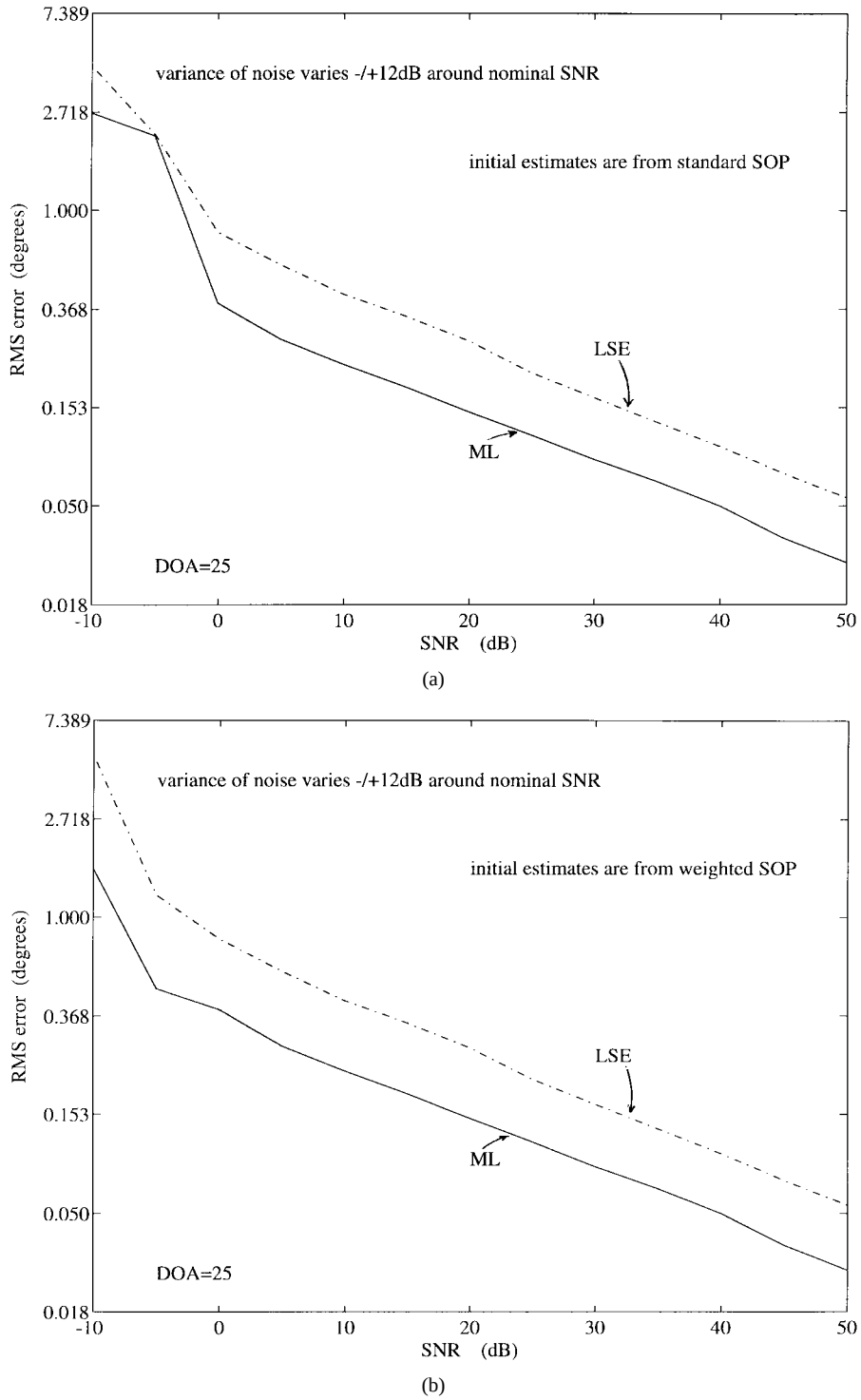

Fig. 5. Statistical comparison of the standard LS algorithm with logarithmic weighting under changing variance of ± 12 dB around nominal SNR. (a) Initial estimates are from sequential orthogonal projection algorithm. (b) Initial estimates are from weighted sequential orthogonal projection algorithm.

III. DOA Estimation and Simulation Results

Let us assume that m narrowband plane waves with center frequency ω_o are incident on an array of p sensors. The $p \times 1$ snapshot vector set $\underline{x}(t_n)$, $n = 1, 2, \cdots, N$ corresponds to samples of the p sensor signals. Vector component $x_k(t_n)$ contains the nth sample of the kth sensor signal. The steering vector $\underline{s}_i(\underline{\theta}_i)$ for the ith plane wave is specified by

$$\underline{s}_i(\underline{\theta}_i) = [e^{-j\omega_o\tau_{i,1}}\, e^{-j\omega_o\tau_{i,2}} \cdots e^{-j\omega_o\tau_{i,p}}]^T \qquad (30)$$

where $\tau_{i,k}$ is the time it takes for the ith plane wave to travel from the kth sensor to the origin. With ν designating the medium's speed of propogation, the time delays $\tau_{i,k}$ can be expressed as

$$\tau_{i,k} = \frac{1}{\nu}[z_k(1)\cos(\theta_i) + z_k(2)\sin(\theta_i)] \qquad \text{for } 1 \le k \le p \quad (31)$$

where the DOA of the ith incident plane wave is designated by θ_i, and $z_k(1)$ and $z_k(2)$ are the x and y coordinates of the kth sensor on the z plane.

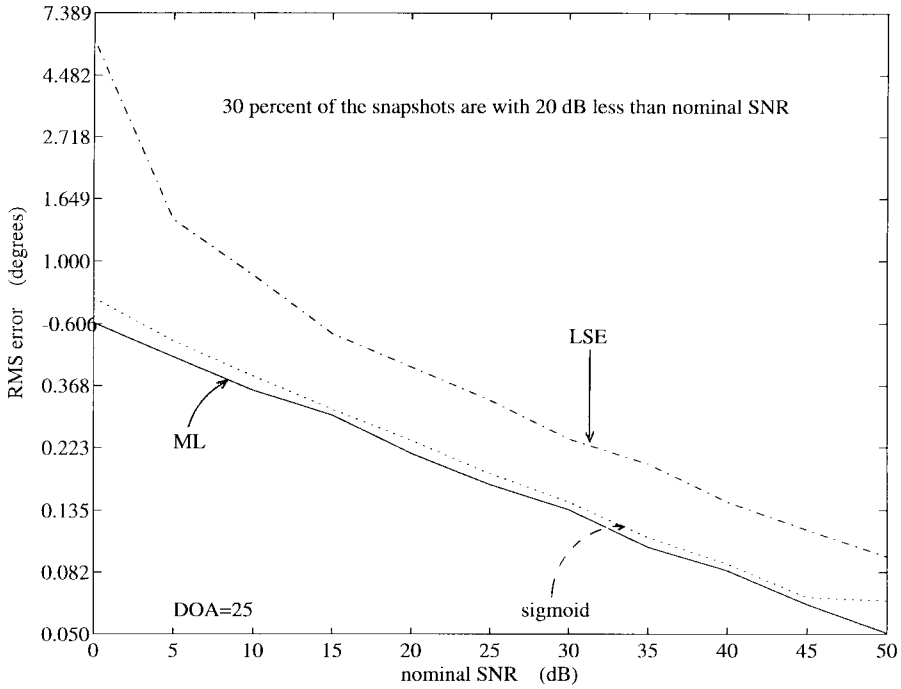

Fig. 6. Statistical comparison of the standard LS algorithm with nonlinear weighting in the presence of data outliers with variance 20 dB less than nominal SNR. $\epsilon = 0.3$. Initial estimates are from weighted SOP algorithm.

Let us now consider the Mills–Cross array, which was used in an experiment between San Diego, CA, and Ottawa, Ont., Canada. This array is composed of 12 sensors ($p = 12$) positioned on the z plane at [−287.5 0], [−187.5 0], [−87.5 0], [87.5 0], [187.5 0], [287.5 0], [0 −287.5], [0 −187.5], [0 −87.5], [0 87.5], [0 187.5], and [0 287.5], where the units are in feet. Two incoherent plane waves with azimuth angles 21° and 25° and a center frequency of 14.85 MHz impinge on this array, as shown in Fig. 3. The speed of propogation is 3×10^8 m/s. The complex valued envelope is generated from a zero-mean unit variance Gaussian process. A total of 40 delayed samples $(N = 40)$ of the input signal are obtained at each sensor.

A. Case I: Changing Variance

In this section, we compare the performance of the standard LS with the ML solution of logarithmic weighting (11) under changing variance conditions. The noise sequence is generated as a p-variate Gaussian distribution $(p = 12)$ with covariance matrix $\sigma_n^2 I_p$. The variance for the nth snapshot σ_n^2 for $1 < n < N$ is obtained so that the signal-to-noise ratio (SNR) takes values uniformly within a range of ±20 dB around a nominal SNR. One hundred trial runs of both methods are performed at "nominal" SNR's starting from 0–50 dB in steps of 5 dB. The Gauss–Newton optimization method in conjunction with Gram–Schmidt orthogonalization is used for optimization.

The effectiveness of the Gauss–Newton descent method is highly dependent on the selection of the initial estimates. When the initial estimates are far from their actual values, the algorithm may converge to a relative minimum. The sequential orthogonal projection (SOP) algorithm, which is also known as the coordinate descent algorithm [20], is shown to be successful for Gaussian noise with constant variance [1], [2]. A weighted version of the SOP algorithm is described in [19]. To test the effectiveness of the initial estimates from the weighted SOP method, the experiments are performed with initial estimates from the weighted and standard SOP methods. For the weighted SOP algorithm, the nonlinear weighting function ψ is chosen as the weighting function corresponding to the log-likelihood function (14). The root-mean-squared (RMS) error versus nominal SNR plots for the plane wave with 25° as the azimuth angle are shown in Fig. 4(a) and (b). For low SNR's, the initial estimates obtained from the standard SOP method are frequently far from the actual DOA's. As a result, the Gauss–Newton method may converge to a relative minimum. When the weighted SOP method [19] is employed, the initial estimates are generally closer to the actual DOA's. Hence, the weighted SOP algorithm provides more reliable initial estimates.

Another set of simulations are performed with noise samples whose variance varies in the range ±12 dB around a nominal SNR. In this case, the noise samples typically do not have as high variances as the previous example for a given nominal SNR. Hence, the effect of weighting in the initial estimates is recognized at lower nominal SNR's. Fig. 5(a) and (b) depicts the performance of the optimally weighted and LSE algorithms with respect to each other when the initial estimates are obtained through the standard and weighted SOP algorithms.

B. Case II: A Spherically Symmetric Distribution

In this case, noise samples are generated from a p-variate Cauchy density function. The scale parameter b in (17) is selected so that $-20 \log b = 5k$ for $k = 3, 4, \cdots, 10$. One hundred trial runs are performed at every value of b with the following selection of nonlinear functions:

1) standard LSE $[\psi(r) = r]$;
2) ML estimator $[\psi(r) = \log(r/b^2 + 1)]$;
3) logarithmic weighting $[\psi(r) = \log(r)]$;
4) a heuristically selected nonlinearity $\psi_o(r) = r/(r^2 + \beta_o)$.

The nonlinear function $\psi_o(r) = r/(r^2 + \beta_o)$ is not monotone increasing but of the redescending type. It is included to add variety. This nonlinearity is expected to limit the effect of the samples at the tails of the Cauchy distribution more than the other three functions. However, the Gauss–Newton method is more likely to converge to a local optimum for this nonlinearity especially when the initial estimates are far from their actual values. At every

iteration of the Gauss–Newton method, the parameter β_o is selected so that the nonlinear function $\psi_o(r)$ takes its maximum value at $r = 2.\text{median}\{r(t_1), r(t_2), \cdots, r(t_N)\}$. In this way, most of the samples are guaranteed to be to the left of its maximum point, and the Gauss–Newton method is less likely to diverge.

The RMS errors of the DOA estimates obtained with the above weighting functions are shown in Table I. It is clear that standard LS methods yield inferior and often unacceptable estimates of the DOA's. The ML weighting scheme performed the best, and the simple logarithmic weighting performed almost identically. The heuristically selected nonlinearity $\psi_o(r)$ yields acceptable results, but it is inferior to the ML and logarithmic weighting.

C. Case III: Outliers in Data

To test the performance of the nonlinear least squares algorithm with different weighting functions, 30% of the 40 snapshots are randomly selected to have an SNR of 20 dB less than their nominal value. The initial estimates are obtained by the weighted SOP technique with optimal weighting of (22). With these initial estimates, the Gauss–Newton method is used to minimize the weighted squared error criterion with weighting function as obtained from

1) the ML weighting function in (22);
2) logarithmic weighting $\psi(r) = \log(r)$;
3) the sigmoid $\psi(r) = a[1 - \exp(-\beta r)]$ with the upper bound as twice the median of the squared error;
4) the standard LS method.

This experiment is repeated 100 times for the same set of nominal SNR's as in Case I. As depicted in Fig. 6, the nonlinear weighting functions yield better results than the standard LS. The performance of the sigmoid is close to that of the ML solution and is a viable alternative in cases where information on the contamination rate and variance are not available. The simple logarithmic weighting again performed similar to the ML weighting function, as mentioned in Section II-C.

The same experiment is repeated for the case in which 10% of the snapshots have a SNR of 20 dB less than nominal SNR. A smaller contamination level ϵ resulted in a smaller gap between RMS errors of the standard LS and the nonlinear weighting. Finally, another experiment with 30% of the snapshots having an SNR of 12 dB less than nominal value is performed. The nonlinear weighting still outperformed the standard LS; however, the difference is not as significant. The advantage of employing a nonlinear function is apparent when either the difference between variances of the nominal and the contaminating distributions is significant, and/or the contamination level ϵ is large.

IV. Conclusions

In this correspondence, a robust DOA estimation method is developed. The robustness is achieved by introducing a nonlinear function that weights the squared error term in the sum-of-squared-error criterion. Weighting functions for various observation noise scenarios, including the Gaussian noise with time-varying variance, the class of spherically symmetric distributions, and ϵ-contaminated Gaussian noise, are determined by the ML estimation theory. It is seen that an appropriately selected nonlinear weighting function improves the estimates of the parameters, and yet, computational complexity of the parameter estimation problem does not increase significantly.

Acknowledgment

The authors would like to thank Associate Editor H. Messer-Yaron and the anonymous reviewers for their positive criticism.

References

[1] J. A. Cadzow, "Least squares error modeling with signal processing applications," *IEEE Acoust., Speech, Signal Processing Mag.*, pp. 12–31, Oct. 1990.
[2] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1553–1560, Oct. 1988.
[3] J. Boheme, "Estimating the source parameters by maximum likelihood and nonlinear regression," in *Proc. ICASSP*, 1984, pp. 7.3.1–7.3.4.
[4] P. J. Huber, *Robust Statistics.* New York: Wiley, 1981.
[5] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey," *Proc. IEEE*, vol. 73, pp. 433–481, Mar. 1985.
[6] J. M. Singer and P. K. Sen, "Asymptotic relative efficiency of multivariate M-estimators," *Commun. Statist.-Simul. Comput.*, vol. 14, no. 1, pp. 29–41, 1985.
[7] S. G. Oh and R. L. Kashyap, "A robust approach for high resolution frequency estimation," *IEEE Trans. Signal Processing*, vol. 39, pp. 627–643, Mar. 1991.
[8] M. Shao and C. L. Nikias, "Signal processing with fractional order moments: Stable processes and their applications," *Proc. IEEE*, vol. 81, pp. 986–1009, July 1993.
[9] H. Messer and P. M. Schultheiss, "On time delay estimation in non-Gaussian noise," in *Proc. IEEE Seventh SP Workshop Statist. Signal Array Process.*, June 1994, pp. 67–70.
[10] G. Veitch and A. R. Wilks, "A characterization of Arctic undersea noise," *J. Acoust. Soc. Amer.*, vol. 77, pp. 989–999, 1985.
[11] M. Bouvet and S. C. Schwartz, "Underwater noises: Statistical modeling, detection, and normalization," *J. Acoust. Soc. Amer.*, vol. 83, no. 3, pp. 1023–1033, 1988.
[12] V. H. Hansen, "Detection performance of some nonparametric rank tests and an application to radar," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 309–318, May 1970.
[13] P. A. Bello and R. Esposito, "A new method for calculating probabilities of errors due to impulsive noise," *IEEE Trans. Commun. Technol.*, vol. COMM-17, pp. 368–379, June 1969.
[14] D. B. Williams and D. H. Johnson, "Robust estimation of structured covariance matrices," *IEEE Trans. Signal Processing*, vol. 41, pp. 2891–2906, Sept. 1993.
[15] D. D. Lee and R. L. Kashyap, "Robust maximum likelihood bearing estimation in contaminated Gaussian noise," *IEEE Trans. Signal Processing*, vol. 40, pp. 1983–1983, Aug. 1992.
[16] G. H. Golub and V. Pereyra, "The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate," *SIAM J. Numer. Anal.*, pp. 413–432, Apr. 1973.
[17] I. Nimmo-Smith, "Linear regressions and sphericity," *Biometrika*, vol. 66, no. 2, pp. 390–392, 1979.
[18] P. Tsakalides and C. L. Nikias, "Maximum likelihood localization of sources in noise modeled as a stable process," *IEEE Trans. Signal Processing*, vol. 45, pp. 2700–2713, Nov. 1995.
[19] Y. Yardımcı, "New results in point source location problem," Ph.D. dissertation, Vanderbilt Univ., Nashville, TN, 1994.
[20] D. G. Luenberger, *Optimization by Vector Space Methods.* New York: Wiley, 1969.