MAXIMUM LIKELIHOOD FOR BLIND SEPARATION AND DECONVOLUTION OF NOISY SIGNALS USING MIXTURE MODELS

Éric Moulines, Jean-François Cardoso, Elisabeth Gassiat

Ecole Nationale Supérieure des Télécommunications (ENST), Département Signal
46, rue Barrault, 75634 Paris Cedex 13, France
Fax: 33 1 45 88 79 35; Email: moulines@sig.enst.fr

ABSTRACT

In this paper, an approximate maximum likelihood method for blind source separation and deconvolution of noisy signal is proposed. This technique relies upon a data augmentation scheme, where the (unobserved) input are viewed as the missing data. In the technique described in this contribution, the input signal distribution is modeled by a mixture of Gaussian distributions, enabling the use of explicit formula for computing the posterior density and conditional expectation and thus avoiding Monte-Carlo integrations. Because this technique is able to capture some salient features of the input signal distribution, it performs generally much better than third-order or fourth-order cumulant based techniques.

1. INTRODUCTION

This contribution is devoted to blind source separation and blind deconvolution. In these models, the observed data $\{x(t)\}$ (a $m \times 1$ vector) is assumed to be given by

$$x(t) = \sum_{k=0}^{L} A(k)s(t-k) + v(t) \qquad (1)$$

where $\{s(t)\}$ $(n \times 1)$ is the (unobserved) input, $\{v(t)\}$ is the (unobserved) additive noise, and $A(z) = \sum_{k=0}^{L} A(k)z^{-k}$ is an unknown $p \times q$ FIR transfer function. It is assumed in the sequel that: **(A1)** $\{v(t)\}$ is an i.i.d additive zero-mean Gaussian noise $v(t) \sim \mathcal{N}(0, J)$ with J a positive definite $m \times m$ matrix **(A2)** $\{s(t)\}$ is an i.i.d sequence of $n \times 1$ random vectors with independent components, with density function $p(s;\gamma)$ given by:

$$p(s;\gamma) = \prod_{i=1}^{n} p_i(s_i;\gamma_i), \ \gamma = [\gamma_1, \cdots, \gamma_n]$$

(A3) $\{s(t)\}$ and $\{v(t)\}$ are independent. This model arises in many signal processing applications, such has sonar array processing or multichannel data communications. A variety of methods and criteria have been proposed in the litterature to solve the problem. Most of these criteria are based on contrast functions based on higher-order statistics (see [11, 10, 6, 7]). In this paper, we concentrate on a maximum likelihood approach. Denote $\eta = (A(0), A(1), \cdots, A(L), J)$ and $\theta = (\eta, \gamma)$. The inference of θ based on $\mathbf{x} = [x(1), x(2), \cdots, x(T)]$ is a standard parametric problem for which maximum likelihood estimator (under appropriate regularity conditions) may shown to be asymptotically efficient. Computation of this estimate is however a very difficult task, because the distribution of the observed data will typically depend in a very intricate manner upon the unknown parameters.

The EM algorithm is ideally suited for such problems [1]. The EM algorithm is an iterative method for finding the mode of the observed (or incomplete) data likelihood $\theta \to p(\mathbf{x};\theta)$, which is extremely useful for many common models for which it is hard to maximize $p(\mathbf{x};\theta)$ directly but easy to work with an extended (complete) observation model, obtained by adding to the observed data appropriately chosen missing data. In blind identification context, the missing data are obviously the input data $s(t)$, which, if available, would make the identification a simple task.

The EM algorithm formalizes a relatively old idea for handling missing data, starting with a guess of the parameters: (i) replace missing values by their expectations given the guessed parameters (ii) estimate parameters assuming the missing data are given by their estimated values, (iii) reestimate the missing values assuming the new parameters are correct, (iv) reestimate parameters and so forth, iterating until convergence. In fact, the EM algorithm is more efficient than these four steps would suggest since each missing data is not estimated separately; instead, those functions of the missing data that are needed to estimate the model parameters are estimated jointly.

The name 'EM' comes from the two alternating steps: finding the *expectation* of the needed functions (sufficient statistics) of the missing values, and *maximizing* to estimate the parameters as if these functions of the missing data were observed. For many standard models, both steps- estimating the missing values gien a current estimate of the parameters and estimating the parameters given the current estimate of the missing values- are straightforward. Unfortunately, this is not the case in the blind identification case: the M-step is more often straightforward (especially when the observation noise is Gaussian; see below), but the E-step is for most input signal distributions analytically untractable (the posterior distribution of the input signal s given x cannot even be expressed in closed form). One solution to this problem is to resort to Monte-Carlo Markov Chain estimation technique (to simulate missing values under the posterior distribution) and to use a stochastic version of the EM algorithm, where basically the expectation step is replaced by Monte-Carlo integration (see, [5]).

It is the purpose of this contribution to show that, by (i) modeling the distribution of the input data as a mixture of Gaussian distributions, and (ii) (in the convolutive mixture case) *splitting* the likelihood function, the E-step can be implemented exactly in a finite (and reasonably small) number of operations, thus avoiding Monte-Carlo integration [3]. We end up with an implementation of the MLE which is efficient both statistically and numerically.

2. NOISY SOURCE SEPARATION AND I/O IDENTIFICATION

We will first concentrate on instantaneous mixture ($L = 0$):

$$x(t) = As(t) + v(t) \tag{2}$$

where A is an unknown $m \times n$ matrix ($m \geq n$). We will then give some hints to extend the results to convolutive mixtures in the next section.

Identifiability in this model is discussed at length in [6]. Basically, the mixing matrix A can be estimated up to (i) permutation of the columns (the numbering of the sources is immaterial) (ii) a global scale factor for each column. To avoid trivialities, we set the diagonal elements of the mixing matrix A to 1. We denote: $\eta = (A, J)$. The problem is to find the value of the parameters $\theta = (\eta, \gamma)$ given $\mathbf{x} = [x(1), \cdots, x(T)]$. The posterior density of $x(t)$ given $s(t) = s$ is $p(x|s; A) = \phi(x; As, J)$ where $\phi(\bullet; m, R)$ denotes the multivariate Gaussian density function with mean μ and covariance matrix R :

$$\log \phi(x; m, R) = -\frac{1}{2}(x - m)' R^{-1}(x - m) - \frac{1}{2} \log |2\pi R|.$$

where $|\bullet|$ denotes the determinant. Denote: $\mathbf{s} = [s(1), s(2), \cdots, s(T)]$. The complete data likelihood is:

$$\log p(\mathbf{x}, \mathbf{s}; \theta) = \sum_{t=1}^{T} \log \phi(x(t); As(t), J) + \sum_{t=1}^{T} \log f(s(t); \gamma)$$

Complete data maximum likelihood estimates for A and J are obtained by maximizing the previous expression under the linear constraints $a_{11} = 1, \cdot, a_{nn} = 1$. This is a quadratic minimization problem under linear constraints, the solution of which problem can be found explicitly (it is not shown here for brevity). The solution depends on the following complete data sufficient statistics :

$$\hat{R}_{xs} = \frac{1}{T} \sum_{t=1}^{T} x(t) s(t)' \quad \hat{R}_{ss} = \frac{1}{T} \sum_{t=1}^{T} s(t) s(t)'.$$

Taking the input data $\{s(t)\}_{t=1,T}$ as the missing data of the EM algorithm, an iteration of the EM procedure consists in **(i)** (E-step) estimating the conditional expectations of $s(t)$ given $x(t)$ and the current fit of the parameter $\theta^{(k)}$ (E-step):

$$\mathrm{E}_{t,\theta^{(k)}}[s(t)] = \int s p(s|x(t); \theta^{(k)}) ds \tag{3}$$

$$\mathrm{E}_{t,\theta^{(k)}}[s(t)s(t)'] = \int s s' \; p(s|x(t); \theta^{(k)}) \; ds.$$

and obtaining the expected values of the complete data sufficient statistics

$$\bar{R}_{xs}^{(k)} = \frac{1}{T} \sum_{t=1}^{T} x(t) \mathrm{E}_{t,\theta^{(k)}}[s(t)]$$

$$\bar{R}_{ss}^{(k)} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{E}_{t,\theta^{(k)}}[s(t)s(t)']$$

and **(ii)** (M-step) maximizing the expected complete data log-likelihood (M-step) under the same lines as above, simply substituting the sufficient statistics $\hat{R}_{xs}$ and $\hat{R}_{ss}$ by their conditional expectations $\bar{R}_{xs}^{(k)}$ and $\bar{R}_{ss}^{(k)}$.

This is indeed in striking similarity with I/O identification. The only difficulty is the evaluation of the above conditional expectations. If the input takes only a finite number of values (which happens to be the case in digital communication context, the input signal being chosen in a finite alphabet) conditional integration reduces to a finite summation [4]; a possible solution is to try to approximate the integrals in Eq. 3 by direct numerical integration or to use a Monte-Carlo stochastic integration technique. The Bayes rule implies that:

$$p(s|x(t); \theta) = \frac{\phi(x(t); As, J) p(s; \gamma)}{\int \phi(x(t); As, J) p(s; \gamma) ds}$$

and thus, despite the components of $s(t)$ are independent, the posterior density $p(s|x(t); \theta)$ does notfactorize as a product of one-dimensional marginal. As a result, in absence of special structure, the evaluation of the posterior density $p(s|x(t); \theta)$ and of the first and second moments of this p.d.f is a formidable task as soon as $n \geq 3$. Perhaps surprisingly, even simulating random variable under the posterior density is difficult: one has to resort to a Monte-Carlo Markov Chain technique, which is a powerful but numerically involved procedure.

It is one of the purposes of this contribution to show that modeling the input as a mixture of Gaussian distributions gives an attractive solution to this problem.

Gaussian mixtures We consider the case where the pdf of s_i (we drop the time index t for simplicity), $1 \leq i \leq n$ is modeled as a finite mixture of Gaussian distributions:

$$f_i(s_i; \gamma_i) = \sum_{j=1}^{q_i} \pi_{ij} \phi(s_i; \mu_{ij}, \sigma_{ij}^2) \quad \gamma_i = (\pi_i, \xi_i),$$

$$\pi_i = [\pi_{i1}, \cdots, \pi_{iq_i}], \xi_i = (\mu_{i1}, \sigma_{i1}^2, \cdots, \mu_{iq_i}, \sigma_{iq_i}^2)$$

When dealing with mixtures, it is convenient to consider that there exists an hidden random variable z_i, taking its values in a finite set $\mathcal{Z}_i = [1, \cdots, q_i]$ with probability $P(z_i = j) = \pi_{ij}$, $1 \leq j \leq q_i$, such that the conditional pdf of s_i given $z_i = j$ is $p(s_i|z_i = j) = \phi(s_i; \mu_{ij}, \sigma_{ij}^2)$. Under (A1)-(A4), the joint distribution of x, the input data s and the label $z = [z_1, z_2, \cdots, z_n]$ has the following nested structure:

$$\begin{aligned} p(x, s, z; \theta) &= p(x|s; \eta) p(s|z; \xi) p(z; \pi), \qquad (4) \\ &= \phi(x; As, J) \phi(s; \mu_z(\xi), \Gamma_z(\xi)) \pi_z, \\ \mu_z(\xi) &= [\mu_{1z_1}, \mu_{2z_2}, \cdots, \mu_{nz_n}], \\ \Gamma_z(\xi) &= \mathrm{diag}[\sigma_{1z_1}^2, \sigma_{2z_2}^2, \cdots, \sigma_{nz_n}^2]. \end{aligned}$$

It is easily seen that the posterior distribution of s given x, z is Gaussian with mean $\alpha_{xz}(\theta)$ and covariance $\Delta_z(\theta)$ respectively given by:

$$\alpha_{xz}(\theta) = \mu_z(\xi) + \Gamma_z(\xi) A' R_z(\theta)(x - A\mu_z(\xi)), \tag{5}$$

$$\Delta_z(\theta) = \Gamma_z(\xi) - \Gamma_z(\xi) A' R_z(\theta) A \Gamma_z(\xi) \tag{6}$$

$$R_z(\theta) = (A\Gamma_z(\xi)A' + J)^{-1}.$$

Similarly, thanks to the decomposition Eq. 4, the posterior distribution of z given x, $\pi_{zx}(\theta) = p(z|x; \theta)$ is:

$$\begin{aligned} \pi_{zx}(\theta) &= \int p(z, s|x; \theta) ds \qquad (7) \\ &\propto \pi_z \int \phi(x; As, J) \phi(s; \mu_z(\xi)\Gamma_z(\xi)) ds \qquad (8) \\ &\propto \frac{\pi_z}{|\Gamma_z||A' J^{-1} A + \Gamma_z(\xi)^{-1}|} \qquad (9) \end{aligned}$$

Using the basic property of conditional probability, we have:

$$p(s,z|x;\theta) = p(s|x,z;\theta)p(z|x;\theta) \tag{10}$$

Hence, the posterior distribution of s given x, $p(s|x;\theta)$ also is a finite mixture of Gaussian distributions,

$$p(s|x;\theta) = \sum_z p(z|x;\theta)\phi(s;\alpha_{xz}(\theta),\Delta_z(\theta)) \tag{11}$$

Note that the number of components of that mixture is equal to the product of the number of components $q = \prod_{i=1}^n q_i$. The computation of the conditional expectations required in Eq. 3 can thus be done in closed form, avoiding Monte-Carlo integration. The complete-data likelihood can now be decomposed as:

$$\begin{aligned}\log p(\mathbf{x},\mathbf{s},\mathbf{z};\theta) &= \sum_{t=1}^T \log\phi(x(t);As(t),J) \\ &+ \sum_{i=1}^n\sum_{j=1}^{q_i}\log\phi(s_i(t);\mu_{ij},\sigma_{ij}^2)\delta\{z_i(t)=j\} \\ &+ \sum_{i=1}^n\sum_{j=1}^{q_i}\log\pi_{ij}\delta\{z_i(t)=j\},\end{aligned}$$

and the EM reestimation functional $Q(\theta|\theta')$ can thus be splitted in three terms, which can be computed in closed-form by applying Eq. 5, Eq. 7, Eq. 10

$$\begin{aligned}Q_1(\theta|\theta') &= \sum_z p(z|x;\theta')\sum_{t=1}^T\int\log\phi(x(t);As,J)\times \\ &\quad\times\phi(s;\alpha_{x(t)z}(\theta),\Delta_{x(t)z}(\theta))ds, \\ Q_2(\theta|\theta') &= \sum_{i=1}^n\sum_{j=1}^{q_i}\int\log\phi(s_i;\mu_{ij},\sigma_{ij}^2)p(s_i,z_i=j|x(t);\theta'), \\ Q_3(\theta|\theta') &= \sum_{i=1}^n\sum_{j=1}^{q_i}\log\pi_{ij}p(z(t)=j|x(t);\theta').\end{aligned}$$

Maximization of $Q_2(\theta|\theta')$ and $Q_3(\theta|\theta')$ is in simple closed-form. Reestimation formula for the parameter π,μ and σ are not given for brevity.

3. CONVOLUTIVE MIXTURES

We now extend the method presented in the previous section to convolutive mixtures. Maximum likelihood estimation for this model is an involved problem because the complete data log-likelihood does not decompose as a sum of 'marginal' components. Up to now, most of the attention has been focused on the case where input signal is a discrete random vector with a finite number of values; in this situation, the observed signal is an Hidden Markov process - HMM - with a finite hidden chain and standard estimation procedure may be applied. Several authors have also derived approximate M.L. estimation procedure for general input signal pdf based on Monte Carlo Markov Chain methods (see for example [3] and the references therein). We will develop in this contribution an alternative strategy, based on the concept of *split data likelihood*, first introduced by Ryden [2] for finite dimensional HMM estimation. Roughly speaking, the procedure amounts to segment the full observation in blocks of size m and to proceed as if the samples in these different blocks are mutually independent. To keep the notations simple, we assume here that $n = 1$ and set $A(0) = 1$. Extensions to the vector case is (at least in theory !) trivial. Let m be an integer, and denote $p_m(x(1),\cdots,x(m);\theta)$ the joint pdf of $\mathbf{x}(u:v) = [x(u),\ldots,x(v)]^T$:

$$\begin{aligned}p_m(\mathbf{x}(1:m);\theta) \propto \quad & \sigma^{-m}\int\prod_{i=1}^m\phi(x_i;\sum_{k=0}^L A(k)s_{i-k},J)\times \\ & \times p(s_{1-L};\gamma)\cdots p(s_m;\gamma)ds_{1-L}\cdots ds_m.\end{aligned}$$

where σ^2 is the noise variance. To estimate the parameter θ', we shall use the contrast function:

$$K(\theta,\theta') = E_\theta\left(\log p_m(\mathbf{x}(1:m);\theta')\right) - E_\theta\left(\log p_m(\mathbf{x}(1:m);\theta')\right) \tag{12}$$

which is the Kullback-Leibler information between the distributions of $(x(1),\ldots,x(m))$ under θ and θ'. Of course, the procedure is not equivalent to maximum likelihood, and we must first check that $\theta' \to K(\theta,\theta')$ is an appropriate contrast function. Using well-known result on Kullback-Leibler distance, $K(\theta,\theta') \geq K(\theta,\theta)$ with equality if and only if:

$$p_m(\mathbf{x}(1:m);\theta) = p_m(\mathbf{x}(1:m);\theta'), \quad a.e. \tag{13}$$

In other word, $K(\theta,\theta')$ is a contrast function iff the parameters are identifiable on a m-dimensional marginal. It can be shown that:

theorem 1 *Assume that there exists $k \geq 3$ such that $E(|s(1)|^k) < \infty$ and $\mathrm{cum}_k(s(1)) \neq 0$. Then, the identifiability condition Eq. 13 is satisfied.*

This result is proved in an extended version of this paper. It is relies upon identifiability conditions obtained by Giannikis and Swami [11]. Now, we shall consider the following contrast process:

$$l_n(\mathbf{x},\theta') = \sum_{s=0}^{[n/m]}\log p_m(\mathbf{x}(sm:sm+m-1);\theta') \tag{14}$$

and then define $\hat{\theta}_n$ as a maximizer of $l_n(\theta')$ over Θ, the set of admissible parameter values. The idea of such a contrast was proposed by Ryden ([2]) for HMM with finite state hidden Markov chain. Under standard regularity conditions, this estimator may be shown to be consistent and asymptotically normal;(this result can be obtained by a straightforward adaptation of proofs of consistency and asymptotic normality proofs of the maximum likelihood estimate in the i.i.d. case).

Remark: Slightly more general contrast can be considered. Indeed, it may be beneficial to overlap the successive blocks by a certain amount, say r, $1 \leq r \leq n$, i.e. to form

$$l_n(\mathbf{x},\theta') = \sum_{s=0}^{[n/r]}\log p_m(\mathbf{x}(sr:sr+m-1);\theta')$$

The estimators obtained by maximizing these contrasts may be shown to be consistent and asymptotically normal. The asymptotic variance of the estimators depends upon r: overlap may improve the variance.

We know focus on implementation issues. Once again, we will use the EM paradigm, and will model the input signal distribution by a finite mixture of Gaussian distributions. Denote $\mathbf{s}(u:v) = [s(u),\cdots,s(v)]^T$ and $\mathbf{z}(u:v) =$

$[z(u), \cdots, z(v)]^T$, where $z(t)$ is the 'label' random variable associated to $s(t)$. The complete data likelihood can be decomposed in a way very similar to Eq. 4

$$p(\mathbf{x}_m, \mathbf{s}_m, \mathbf{z}_m; \theta) = \sigma^{-m} \times \quad (15)$$
$$\prod_{k=0}^{m-1} \exp(-\frac{1}{2\sigma^2}(x(t+k) - A\mathbf{s}(t-L:t+m-1))^2) \times$$
$$\prod_{k=-L}^{m} \exp(-\frac{1}{\sigma_z^2}(s(t)-\mu_z)^2)\pi_z$$

where $[\mu_1, \cdots, \mu_q]$ and $[\sigma_1^2, \cdots, \sigma_q^2]$ are respectively the mean and the variance of the individual components of the mixture and $\pi = [\pi_1, \cdots, \pi_q]$ are the proportions and where $\mathbf{x}_m = \mathbf{x}(t : t+m-1)$, $\mathbf{s}_m = \mathbf{s}(t-L : t+m-1)$, $\mathbf{z}_m = \mathbf{z}(t-L : t+m-1)$. As above we need to compute the following quantities: $p(\mathbf{s}_m|\mathbf{x}_m, \mathbf{z}_m; \theta)$ and $p(\mathbf{z}_m|\mathbf{x}_m; \theta)$.The conditional distribution $\mathbf{s}_m$ given $\mathbf{x}_m$ and $\mathbf{z}_m$ is Gaussian: the mean and the covariance of this distribution can be obtained very efficiently using an efficient 'Kalman-like' state smoother (while the Kalman filter uses past observations, a state smoother yields estimators of the state vector based on the full set of observations in the sample. The classical fixed interval smoother is presented in [13] but a more efficient state smoother has recently been developed in [8, 9]). Similarly, the conditional distribution of $\mathbf{z}_m$ given $\mathbf{x}_m$ may be seen to be proportional to :

$$p(\mathbf{z}_m|\mathbf{x}_m; \theta) \propto \int p(\mathbf{x}_m, \mathbf{s}_m, \mathbf{z}_m; \theta) d\mathbf{s}_m$$

this integral can be explicitely computed based on the decomposition Eq. 15. Significant reduction of the computational burden can be achieved by resorting to a kind of 'forward-backward' algorithm, coupling these two steps. This algorithm is described in the full-length version of the paper. The remaining steps closely follow those presented in the previous section and are not repeated. This method is computationally intensive but still is an order of magnitude faster than Monte-Carlo based solutions to this problem.

4. SIMULATION

As an illustrative example suppose that the input signal is an i.i.d sequence distributed according to a mixture of two Gaussian distribution:

$$p(s;\gamma) = \pi\phi(s;\mu_1,\sigma_1^2) + (1-\pi)\phi(s;\mu_2,\sigma_2^2)$$

we set: $\pi = 0.5$, $\mu_1 = -\mu_2 = 0.95$ and $\sigma_1^2 = \sigma_2^2 = 0.1$. It is easily seen that: $E_\gamma(s) = 0$ and $E_\gamma(s^2) \approx 1$. The skewness of s is null and the kurtosis is $k_4(s) \approx -1.6$. This signal is fed in a FIR filter of order 3, with coefficients [0.5 -0.75 0.5]. The additive noise $v(k)$ is zero-mean, Gaussian, and the signal to noise ratio is varied between 0 dB and 20 dB. The record size is $T = 200$. We use the splitting procedure described above with $m = 4$. To reduce realization dependency of the simulations, we averaged over 100 Monte-Carlo simulations. The initial estimates for the filter coefficients and the noise level are obtained by applying the Giannikis-Swami [11] fourth-order identification technique. Results are summarized in the table below. It is seen that the procedure produces very reliable estimates of the FIR coefficients. This method clearly outperforms higher-order methods, at the expense of a significant increase in the computational burden.

True MA coef.	0.5	-0.75	0.5
0 dB			
mean	0.52	-0.77	0.47
std. dev.	0.25	0.33	0.23
10 dB			
mean	0.51	-0.74	0.49
std. dev.	0.06	0.11	0.09
20 dB			
mean	0.50	-0.75	0.50
std. dev.	0.01	0.01	0.02

In fact, the proposed method extends to the semi-parametric context, in which the input signal distribution is not known: it is not necessary that the input data actually are distributed as Gaussian mixtures. The crucial point is that the mixture captures the most important features of the distribution of the input. Since many distributions can be approximated with arbitrary precision as Gaussian mixtures, the present approach offers a route to semi-parametric estimation. Practical and theoretical results in that direction will be presented at the conference.

REFERENCES

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39:1–38, 1977.

[2] Ryden, T. (1994). Consistent and asymptotically normal parameter estimates for hidden Markov models. *Annals of Stat.* **22** 1884-1895.

[3] A. Doucet and P. Duvaut. Fully bayesian analysis of conditionally linear gaussian state space model. In *Proc. ICASSP Atlanta*, volume 5, pages 2948–2951, 1996.

[4] A. Belouchrani and J-F Cardoso. Maximum likelihood source separation for discrete sources. In *Proc. EUSIPCO*, pages 768–771, Edinburgh, September 1994.

[5] M. Tanner *Tools for statistical inference Springer Series in Statistics, Springer-Verlag*, 1993

[6] P. Comon Independent component analysis, a new concept ? *Signal Processing*, volume 36, pp. 287-314, 1994

[7] O. Shalvi and E. Weinstein System identification based on higher-order statistics submitted to *IEEE Trans. on Signal Proc.*, 1995

[8] P. de Jong. Smoothing and interpolation with the state space model *J. Am. Statist. Assoc.*, volume 84, 1085-88, 1991.

[9] S. J. Koopman Disturbance smoother for state space models *Biometrika*, volume 80, 117-126, 1993.

[10] J. Tugnait and U. Gummadavelli, Blind channel estimation and deconvolution in colored noise using higher-order cumulants In *Proc. SPIE*, volume 2296, 107-125

[11] G. Giannikis and A. Swami, On estimating noncausal nonminimum phase ARMA models of non-gaussian processes *IEEE Trans. on Acoust. Speech and Sig. Proc.*, volume 38, 478-495, 1990

[12] D. Yellin and E. Weinstein, Multichannel signal separation: methods and analysis, *IEEE Trans. on Acoust. Speech and Sig. Proc.*, volume 44, 106-118, 1996.

[13] B.D.O. Anderson and J.B. Moore *Optimal filtering Englewood cliffs, New Jersey: Prentice-Hall*, 1979