Estimator for Number of Sources using Minimum Description Length Criterion for Blind Sparse Source Mixtures

Radu Balan

Siemens Corporate Research
755 College Road East
Princeton, NJ 08540
radu.balan@siemens.com

Abstract. In this paper I present a Minimum Description Length Estimator for number of sources in an anechoic mixture of sparse signals. The criterion is roughly equal to the sum of negative normalized maximum log-likelihood and the logarithm of number of sources. Numerical evidence supports this approach and compares favorabily to both the Akaike (AIC) and Bayesian (BIC) Information Criteria.

1 Signal and Mixing Models

Consider the following model in time domain:

$$x_d(t) = \sum_{l=1}^{L} s_l(t-(d-1)\tau_l) + n_d(t) \quad , \quad 1 \leq d \leq D \tag{1}$$

This model corresponds to an anechoic Uniform Linear Array (ULA) with L souces and D sensors. In frequency domain, (1) becomes

$$X_d(k,\omega) = \sum_{l=1}^{L} e^{-i\omega(d-1)\tau_l} S_l(k,\omega) + N_d(k,\omega) \tag{2}$$

We use the following notations: $\mathbf{X}(k,\omega)$ for the D-complex vector of components $(X_d(k,\omega))_d$, $\mathbf{S}(k,\omega)$ for the L-complex valued vector of components $(S_l(k,\omega))_l$, and $A(\omega)$ the $D \times L$ complex matrix whose (d,l) entry is $A_{d,l}(\omega) = e^{-i\omega(d-1)\tau_l}$.

In this paper I make the following statistics assumptions:

1. (H1) Noise signals $(n_d)_{1 \leq d \leq D}$ are Gaussian i.i.d. with zero mean and unknown variance σ^2;
2. (H2) Source Signals are unknown, but for every time-frequency point (k,ω), at most one signal $S_l(k,\omega)$ is nonzero, among the total of L signals;
3. (H3) The number of source signals L is a random variable.

The probem is to design a statistically principled estimator for L, the number of source signals. In this paper I study the Minimum Description Length approach for this problem.

For this model, the measured data is $\Xi = \{(X_d(k,\omega))_{1\leq d\leq D}\ ,\ 1 \leq k \leq T, 1 \leq \omega \leq F\}$. Furthermore the number of sensors D is also known. The rest of parameters are unknown. I denote $\theta = (\theta', L)$, where:

$$\theta' = \left(\{(S_l(k,\omega))_{1\leq l\leq L}\ ;\ 1 \leq k \leq T, 1 \leq \omega \leq F\}\ ,\ (\tau_l)_{1\leq l\leq L}\ ,\ \sigma^2\right) \tag{3}$$

Notice that hypothesis (H2) above imposes a constraint on set $(S_l(k,\omega))_{1\leq l\leq L}$, for every (k,ω). More specifically, the L complex vector $(S_l(k,\omega))_{1\leq l\leq L}$ has to lay in one of the L 1-dimensional coordinate axes (that is, all but one component has to vanish). This fact has a profound implication on estimating the complexity penalty associated to the parameters set. Some real world signals may satisfy (H2) only approximately. For instance [1] studies this assumption for speech signals.

1.1 Prior Works

The signal and mixing model described before has been analyzed by many works before.

In the past series of papers [2–7] the authors studied (1), and several generalizations of this model in the following respects. Mixing model: each channel may have an attenuation factor (equivalently, τ_l may be complex); Noise statistics: noise signals may have inter-sensor correlations; Signals: more signals may non-vanish at each time-frequency point (maximum number allowed is $D-1$); more recently we have considered temporal, and time-frequency, dependencies on signal statistics.

A similar model, and a similar sparsness assumption, has been used by the DUET algorithm [1], or by [8], [9].

Similar assumptions to [5] have been made by [10] for an instantaneous mixing model. As the authors mentioned there, as well in [11, 12], and several others, a new signal separation class is defined by sparsness assumption, called Sparse Component Analysis (SCA). In this vein, this present paper proposes a look at the Minimum Description Length paradigm in the context of Sparse Component Analysis.

Before discussing the new results of this paper, I would like to comment on other approaches to the BSS problem. Many other works dealt with the mixing model (1), or its generalizations to a fully echoic model. A completely different class of algorithms is furnished by the observation that, in frequency domain, the echoic model simply becomes an instantaneous mixing model. Therefore standard ICA techniques can be applied, as in [13, 14] to name a few. Next, one has to connect frequency domain components together for the same source. The permutation ambiguity is the main stumbling block. Several approaches have been proposed, some based on ad-hoc arguments, [15, 9]. A more statistically principled approach has been proposed and used by Zibulevsky [16] and in more recent papers, as well as by other authors, by assuming a stochastic prior model for source signals. The Maximum A Posteriori (MAP), or Minimum Mean Square Error (MMSE) estimators can be derived. While principly they are superior to Maximum Likelihood type estimators derived in [4, 5], or mixed estimators such as [1, 8, 9], they require a good prior stochastic model. This makes difficult the comparison between classes of BSS solutions.

In the absence of noise, the number of sources can be estimated straightforwardly by building a histogram of the instantaneous delay (τ), or for a more general model see [10].

As I mention later, the MDL paradigm here may be well applied in conjunction with other signal estimators, in particular with the MAP estimators described before.

2 Estimators

Assume the mixing model (1) and hypotheses (H1),(H2),(H3). Then its associated likelihood is given by

$$\mathcal{L}(\theta) := P(\Xi|\theta) = \prod_{(k,\omega)} \frac{1}{\pi^D \sigma^{2D}} exp\left(-\frac{1}{\sigma^2}\|\mathbf{X}(k,\omega) - A(\omega)\mathbf{S}(k,\omega)\|^2\right) \tag{4}$$

In the next subsection the maximum likelihood estimator for θ', and the maximum likelihood value are going to be derived.

Following a long tradition of statistics papers, consider the following framework. Let $P(X)$ denote the unknown true probability of data (measurements), $P(X|\theta)$ denote the data likelihood given the model (1) and (H1-H3). Then the estimation objective is to minimize the misfit between these two distributions measured by a distance between the two distribution functions. One can choose the Kullback-Leibler divergence, and obtain the following optimization criterion:

$$J(\theta) = D(P_X||P_{X|\theta}) := \int log\frac{P(X)}{P(X|\theta)}dP(X) = \int \log\, P(X)\, dP(X) - \int \log\, P(X|\theta)\, dP(X) \tag{5}$$

Since the first term does not depend on θ, the objective becomes maximization of the second term:

$$\hat{\theta} = argmax_\theta \mathbf{E}[\log\, P_{X|\theta}(X|\theta)] \tag{6}$$

where the expectation is computed over the true data distribution P_X. However the true distribution is unknown. A first approximation is to replace the expectation $\mathbf{E}$ by average over data points. Thus one obtains the maximum likelihood estimator (MLE):

$$\hat{\theta}_{ML} = argmax_\theta \frac{1}{N}\sum_{t=1}^{N} \log\, P_{X|\theta}(X_t|\theta) \tag{7}$$

where N is the number of sample points $(X_t)_{1\leq t\leq N}$.

As is well known in statistical estimation (see [17, 18]), the MLE is usually biased. For discrete parameters, such as number of source signals, this bias has a bootstraping effect that monotonically increases the likelihood and makes the number of parameter estimation impossible through naive MLE. Several approaches proposed to estimate and make correction for this bias. In general, the optimization problem is restated as:

$$\hat{\theta} = argmin_\theta \left[-\frac{1}{N}\sum_{t=1}^{N} \log\, P(X_t|\theta) + \Phi(\theta, N)\right] \tag{8}$$

Following e.g. [18] we call Φ the *regret.* Akaike [17] proposes the following regret:

$$\Phi_{AIC}(\theta,N) = \frac{|\theta|_0}{N} \tag{9}$$

where $|\theta|_0$ represents the total number of parameters. Schwarz [19] proposes a different regret, namely

$$\Phi_{BIC}(\theta,N) = \frac{|\theta|_0 \log N}{2N} \tag{10}$$

In a statistically plausible interpretation of the world, Rissanen [20] obtains for regret the shortest possible description of the model using the universal distribution function of Kolmogorov, hence the name *Minimum Description Length*,

$$\Phi_{MDL}(\theta,N) = Coding\ Length_{Kolmogorov\ p.d.f.}(Model(\theta,N)) \tag{11}$$

Based on this interpretation, $\Phi(\theta,N)$ represents a measure of the model complexity.

My approach here is the following. I propose the following regret function

$$\Phi_{MDL-BSS}(\theta,N) = log_2(L) + \frac{L\,log_2(M)}{N} \tag{12}$$

where M represents precision in optimization estimation of delay parameters τ (for instance the number of grid points of an 1-D exhaustive search). Thus the optimization in (8) is carried out in two steps. First, for fixed L, the log likelihood is optimized over θ':

$$\hat{\theta}'_{MLE}(L) = argmax_{\theta'} P(X|\theta',L) \quad , \quad MLV(L) = P(X|\hat{\theta}'_{MLE},L) \tag{13}$$

Here MLV denotes the Maximum Likelihood Value. Then L is estimated via:

$$\hat{L}_{MDL-BSS} = armin_L \left[-\log(MLV(L)) \ + \ \log_2(L) \ + \ \frac{L\,log_2(M)}{N} \right] \tag{14}$$

In the next subsection I present the computation of the Maximum Likelihood Value (MLV). Then, in the following subsection I argue the particular form (12) for $\Phi(\theta,N)$ inspired by the MDL interpretation. In same subsection I also present difficulties in a straightforward application of AIC or BIC criteria.

2.1 The Maximum Likelihood Value

The material from this subsection is presented in more detail in [4]. Results are summarized here for the benefit of the reader.

The constraint (H2) assumed in section 1 can be recast by introducing the selection variable $V(k,\omega)$: $V(k,\omega) = l$ iff $S_l(k,\omega) \neq 0$, and the complex amplitudes $G(k,\omega)$. Thus a slightly different parametrization of the model is obtained. The new set of parameters is now $\psi = (\psi', L)$ where

$$\psi' = \left(\{(G(k,\omega), V(k,\omega)) \ ; \ 1 \leq k \leq T, 1 \leq \omega \leq F\} \ , \ (\tau_d)_{1\leq d \leq D} \ , \ \sigma^2 \right) \tag{15}$$

The signals in θ' are simply obtained through: $S_{V(k,\omega)}(k,\omega) = G(k,\omega)$, and $S_l(k,\omega) = 0$ for $l \neq V(k,\omega)$.

The likelihood (4) becomes:

$$\mathcal{L}(\psi) = \frac{1}{\pi^{DN}\sigma^{2DN}} exp\left(-\frac{1}{\sigma^2}\sum_{(k,\omega)} \|\mathbf{X}(k,\omega) - G(k,\omega)A_{V(k,\omega)}(\omega)\|^2\right) \tag{16}$$

where N is the number of time-frequency data points, and $A_l(\omega)$ denotes the l^{th} column of matrix $A(\omega)$. The optimization over G is performed immediately, as a least square problem. The optimum value is replaced in $\mathcal{L}(\psi)$:

$$log\mathcal{L}((V)_{k,\omega},(\tau_l)_l,L) = -DN\,log(\pi) - DN\,log(\sigma^2) - \frac{1}{\sigma^2}\sum_{k,\omega}\left[\|\mathbf{X}(k,\omega)\|^2 - \frac{1}{D}|\langle\mathbf{X}(k,\omega),A_{V(k,\omega)}(\omega)\rangle|^2\right]$$

The optimization over $(V)_{k,\omega}$ and $(\tau_l)_{1\le l\le L}$ is performed iteratively as in the K-means algorithm:

- For a fixed set of delays $(\tau_l)_l$, the optimal selection variables are
$$V(k,\omega) = argmax_m|\langle\mathbf{X}(k,\omega),A_m(\omega)\rangle| \tag{17}$$
- For a fixed selection map $(V(k,\omega))_{k,\omega}$, consider the induced partition $\Pi_m = \{(k,\omega)\ ;\ V(k,\omega) = m\}$. Then τ_m is obtained by solving L 1-dimensional optimization problems
$$\tau_m = argmax_\tau \sum_{(k,\omega)\in\Pi_m} |\langle\mathbf{X}(k,\omega),A_m(\omega;\tau)\rangle|^2 \tag{18}$$

This steps are iterated until convergence is reached (usually is a relatively small number of steps, e.g. 10). Denote $\hat{V}_{MLE}(k,\omega)$ and $\hat{\tau}_{lMLE}$ the final values, and replace these values into $\mathcal{L}$. The noise variance parameter is estimated by maximizing $\mathcal{L}$ over σ^2,

$$\hat{\sigma^2}_{MLE} = \frac{1}{N}\sum_{(k,\omega)}\left[\|\mathbf{X}(k,\omega)\|^2 - \frac{1}{D}|\langle\mathbf{X}(k,\omega),A_{\hat{V}_{MLE}(k,\omega)}(\omega;\hat{\tau}_{MLE}\rangle|^2\right] \tag{19}$$

Finally, the log maximum likelihood value becomes:

$$log(MLV(L)) = \frac{1}{N}log(\mathcal{L}(\hat{\psi'}_{MLE};L)) = -D\,log(\pi) - 1 - D\,log(\hat{\sigma^2}_{MLE}) \tag{20}$$

where $\hat{\psi'}_{MLE}$ denoted the optimal parameter set ψ' containing the combined optimal values $(\hat{V}_{MLE}(k,\omega))_{(k,\omega)}$, $(\hat{G}_{MLE}(k,\omega))_{(k,\omega)}$, $(\hat{\tau}_l)_{1\le l\le L}$, $\hat{\sigma^2}_{MLE}$.

2.2 Number of Sources Estimation

The next step is to establish the regret function. As mentioned earlier the approach here is to use an estimate of the Minimum Description Length of the model (1) together with hypotheses (H1-H3). In general this is an impossible task since the Kolmogorov's universal distribution function is unkown. However the L-dependent part of the model description is embodied in the mixing parameters $(\tau_l)_{1\le l\le L}$, and the selection map $(V(k,\omega))_{(k,\omega)}$. Approximating by a uniform distribution in the space of delays with a finite discretization of, say,

M levels, and no prior preferential treatment of one source signal versus the others, an upper bound on the description length is obtained as the code length of an entropic encoder for this data added to the description length of the entire sequence of models with respect to the Kolmogorov universal distribution:

$$l^*(Model;N) \leq Llog_2(M) + Nlog_2(L) + C(Model) \tag{21}$$

This represents an upper bound since $l^*(Model;N)$ is supposed to represent the optimal description (minimal description) length, whereas the description splits into two parts: the sequence of models parametrized by ψ and N, and then, for a given (L,N) the entropic length of ψ. This clearly represents only one possible way of encoding the pair $(Model(\psi),N)$.

This discussion justifies the following choice for the regret function $\Phi_{MDL-BSS}$

$$\Phi_{MDL-BSS}(L,N) = \frac{L\,log_2(M) + N\,log_2(L)}{N} = log_2(L) + \frac{L\,log_2(M)}{N} \tag{22}$$

as mentioned earlier in (12).

Before presenting experimental evidence supporting this approach, I would like to comment on AIC and BIC criteria. The main difficulty comes from the estimation of the number of parameters. Notice that, using θ description, the number of parameters becomes $LN+L+2$, whereas in ψ description, this number is only $2N+L+2$. The difference is due to that fact that the set of realizable signal vectors $(S_l)_{1\leq l\leq L}$ lays in a collection of L 1-dimensional spaces. Thus this can be either modeled as a collection of L variables, or by 2 variables: complex amplitude, and a selection map V. Consequently, the regret function for AIC can be either $L+\frac{L+2}{N}$, or $2+\frac{L+2}{N}$. Similarly, for BIC the regret function can be $L\,log(N)/2+\frac{(L+2)log(N)}{2N}$, or $log(N)+\frac{(L+2)log(N)}{2N}$. The criterion I propose in (22) interpolates between these two extrema, and, in my opinion, it captures better the actual size of model parametrization.

3 Experimental Evaluation

Consider the following setup. A Uniform Linear Array (ULA) with a variable number of sensors runging from 2 to 5, and distance between adjacent sensors of 5 cm, that records anechoic mixtures of signals coming from $L \leq 6$ sources. The sources are spread uniformly with a minimum of 30 degrees separation. Additive Gaussian noise of average SNR ranging from 10dB to 100dB has been added to recordings. The signals were TIMIT voices sampled at 16 KHz, and each of length 38000 samples (roughly 3 male and female voices saying "She had a dark suit in a greasy wash water all year").

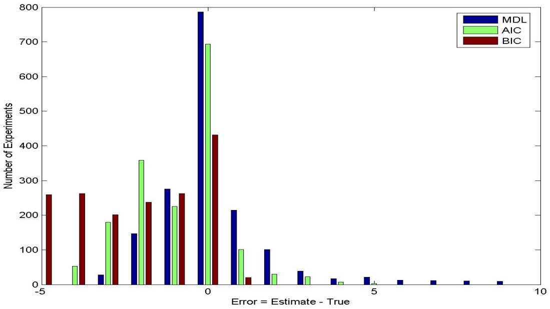
For this setup, the noise was varied in 10dB steps, and number of sources ranged from 1 to 6. The delay optimization (18) was performed through a grid search with step 0.05 samples. Since $\tau_{max} = 2.4$, there were $M = 96$ possible values of τ. Thus $\frac{log_2(M)}{N} = 1.7\,10^{-4}$ and the correction term $L\frac{log_2(M)}{N}$ in $\Phi_{MDL-BSS}$ had no influence. Similarly, the $\frac{L}{N}$ term in AIC and $\frac{L\,log(N)}{N} = 3\,10^{-4}\,L$ in BIC are too small. Therefore the only meaningful AIC and BIC were given by the former regret functions. To summarize, the source number estimator is given by:

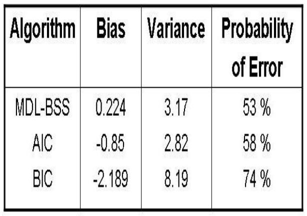
$$\hat{L}_{MDL-BSS} = argmin_L \left[-log\, MLV(L) + log_2(L)\right] \tag{23}$$

$$\hat{L}_{AIC} = argmin_L \left[-log\, MLV(L) + L\right] \tag{24}$$

$$\hat{L}_{BIC} = argmin_L \left[-log\, MLV(L) + L\, log(N)\right] \tag{25}$$

where the optimization is done by exhaustive search for L over the range 1 to 10. For a total of 1680 experiments (10 levels of noise x 4 number of sensors x 6 number of sources x 7 realizations), the histogram of estimation error has been obtained. For each of the three estimators, the histogram is rendered in Figure 1. Statistical performance of these estimators is presented in Table at right.





Algorithm	Bias	Variance	Probability of Error
MDL-BSS	0.224	3.17	53 %
AIC	-0.85	2.82	58 %
BIC	-2.189	8.19	74 %

Fig. 1. The histograms of estimation errors for MDL-BSS criterion (left bar), AIC criterion (middle bar), BIC criterion (right bar). Table with statistical performance of the three estimators.

4 Conclusions

The MDL-BSS estimator clearly performed best among the three estimators, since the error distribution is the most concentrated to zero, in every sense: the number of errors is the smallest, the average error is the smallest, the variance is the smallest, the bias is the smallest. Estimation error is explained by a combination of two factors: 1) source signals (voices) do not satisfy the hypothesis (H2), instead there is always an overlap between time-frequency signal supports; and 2) the estimates for location, noise variance, and separated signals were biased; this bias compounded and inverted the minimum position. The other two estimators (AIC, and BIC) were biased towards underestimating the number of sources.

This paper provides a solid theoretical footing for a statistical criterion to estimate number of source signals in an anechoic BSS scenario with sparse signals. Extension to other mixing models (such as instantaneous) is obvious. The regret function stays the same, only the MLV is modified. The same approach can be used to other Sparse Component Analysis, and this analysis will be done elsewhere.

The numerical simulations confirmed the estimation performance.

References

1. O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. on Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, July 2004.
2. S. Rickard, R. Balan, and J. Rosca, “Real-time time-frequency based blind source separation,” in *Proc. ICA*, 2001, pp. 651–656.
3. R. Balan, J. Rosca, and S. Rickard, “Non-square blind source separation under coherent noise by beamforming and time-frequency masking,” in *Proc. ICA*, 2003.
4. R. Balan, J. Rosca, and S. Rickard, “Scalable non-square blind source separation in the presence of noise,” in *ICASSP2003, Hong-Kong, China*, April 2003.
5. J. Rosca, C. Borss, and R. Balan, “Generalized sparse signal mixing model and application to noisy blind source separation,” in *Proc. ICASSP*, 2004.
6. R. Balan and J. Rosca, “Convolutive demixing with sparse discrete prior models for markov sources,” in *Proc. BSS-ICA*, 2006.
7. R. Balan and J. Rosca, “Map source separation using belief propagation networks,” in *Proc. ASILOMAR*, 2006.
8. M. Aoki, M. Okamoto, S. Aoki, and H. Matsui, “Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones,” *Acoust. Sci. & Tech.*, vol. 22, no. 2, pp. 149–157, 2001.
9. H. Sawada, R. Mukai, S. Araki, and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *IEEE Trans. SAP*, vol. 12, no. 5, pp. 530–538, 2004.
10. P. Georgiev, F. Theis, and A. Cichocki, “Sparse component analysis and blind source separation of underdetermined mixtures,” *IEEE Tran. Neur.Net.*, vol. 16, no. 4, pp. 992–996, 2005.
11. A. Cichocki, Y. Li, P. Georgiev, and S.-I. Amari, “Beyond ica: Robust sparse signal representations,” in *IEEE ISCAS Proc.*, 2004, pp. 684–687.
12. A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, Wiley, April 2002.
13. Pierre Comon, “Independent component analysis, a new concept?,” *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
14. A.J. Bell and T.J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation*, vol. 7, pp. 1129–1159, 1995.
15. J.Annemuller and B.Kollmeier, “Amplitude modulation decorrelation for convolutive blind source separation,” in *ICA*, 2000, pp. 215–220.
16. P. Bofill and M. Zibulevsky, “Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform,” in *Proc ICA*, Helsinki, Finland, June 19–22 2000, pp. 87–92.
17. H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Aut. Cont.*, vol. 19, no. 6, pp. 716–723, 1974.
18. A. Barron, J. Rissanen, and B. Yu, “The minimum description length principle in coding and modeling,” *IEEE Trans. Inf. Th.*, vol. 44, no. 6, pp. 2743–2760, 1998.
19. G. Schwarz, “Estimating the dimension of a model,” *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
20. J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, pp. 465–471, 1978.