Information Theory Based Estimator of the Number of Sources in a Sparse Linear Mixing Model

Radu Balan
University of Maryland
Department of Mathematics, Center for Scientific Computation
And Mathematical Modeling, and Norbert Wiener Center
College Park, MD 20850
rvbalan@math.umd.edu

***Abstract*—In this paper we present an Information Theoretic Estimator for the number of sources mutually disjoint in a linear mixing model. The approach follows the Minimum Description Length prescription and is roughly equal to the sum of negative normalized maximum log-likelihood and the logarithm of number of sources. Preliminary numerical evidence supports this approach and compares favorabily to both the Akaike (AIC) and Bayesian (BIC) Information Criteria.**

I. The Mixing Model and Signals

Consider the following mixing model:

$$x_d(t) = \sum_{l=1}^{L} a_{d,l} s_l(t) + n_d(t)\ ,\ 1 \le d \le D\ ,\ 1 \le t \le T \tag{1}$$

This model corresponds to an Instantaneous Linear Mixing Model with L souces and D sensors. We will frequently use the vector notation $\mathbf{X}(t) = (x_1(t), \ldots, x_D(t))^T$, and matrix $A = (a_{d,l})$.

In this paper the following assumptions are made:

1) (H1) Noise signals $(n_d)_{1\le d\le D}$ are Gaussian i.i.d. with zero mean and unknown variance σ^2;
2) (H2) Source Signals are unknown, but at every moment t at most one signal $s_l(t)$ is nonzero, among the total of L signals;
3) (H3) The number of source signals L is an unknown random variable;
4) (H4) The columns of the unknown matrix A have unit Euclidian norm.

The probem is to design a statistically principled estimator for L, the number of source signals.

For this model, the measured data is $\Xi = \{(x_d(t))_{1\le d\le D}\ ,\ 1 \le t \le T\}$. Furthermore the number of sensors D is also known. The rest of parameters are unknown. Let us denote by $\theta = (\theta', L)$, where:

$$\theta' = \left(\{(s_l(t))_{1\le l\le L}\ ;\ 1 \le t \le T\}\ ,\ \sigma^2\right) \tag{2}$$

Notice that hypothesis (H2) above imposes a constraint on set $(s_l(t))_{1\le l\le L}$, for every t. More specifically, the L-dimensional vector $(s_l(t))_{1\le l\le L}$ 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. We will comment towards the end of this paper on how to extend this hypothesis to the case when M components of $(s_l(t))_{1\le l\le L}$ are allowed to be non-zero, for every t.

A. Prior Works

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

A similar mixing model to (1) has been studied in [1], [2], and [3]. As the authors mentioned there, as well as in [4], [5], 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 framework. The reader is refered to our recent paper [6] where we analyzed similar estimators for number of sources but for an anechoic mixing model.

In the past series of papers [7], [8], [9], [10], [11] the authors studied the non-instantaneous anechoic version of (1), and several generalizations of this

model in the following respects. Mixing model: each channel may have a delay and an attenuation factor; 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 [12].

In the absence of noise, the number of sources can be estimated straightforwardly by building a histogram of the ratios $x_l(t)/x_1(t)$, or for a more general model see [3].

II. Estimators

Assume the mixing model (1) and hypotheses (H1),(H2),(H3),(H4). Then its associated likelihood is given by

$$\begin{aligned} \mathcal{L}(\theta) &:= P(\Xi|\theta) \\ &= \prod_{t=1}^{T} \frac{1}{(2\pi)^{D/2}\sigma^D} exp\left(-\frac{1}{2\sigma^2}\|\mathbf{X}(t)-A\mathbf{S}(t)\|^2\right) \end{aligned} \tag{3}$$

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-H4). 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:

$$\begin{aligned} J(\theta) &= D(P_X||P_{X|\theta}) := \int log\frac{P(X)}{P(X|\theta)} dP(X) \\ &= \int \log\, P(X)\; dP(X) - \int \log\, P(X|\theta)\; dP(X) \end{aligned} \tag{4}$$

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

$$\hat{\theta} = argmax_\theta \mathbf{E}[\log\, P_{X|\theta}(X|\theta)] \tag{5}$$

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):

$$\hat{\theta}_{ML} = argmax_\theta \frac{1}{T}\sum_{t=1}^{T} \log\, P_{X|\theta}(X_t|\theta) \tag{6}$$

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

As is well known in statistical estimation (see [13], [14]), for finite sample size 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:

$$\hat{\theta} = argmin_\theta \left[-\frac{1}{T}\sum_{t=1}^{T} \log\, P(X_t|\theta) + \Phi(\theta, T)\right] \tag{7}$$

Following e.g. [14] we call Φ the *regret*. Akaike [13] proposes the following regret:

$$\Phi_{AIC}(\theta, T) = \frac{|\theta|_0}{T} \tag{8}$$

where $|\theta|_0$ represents the total number of parameters. Schwarz [15] proposes a different regret, namely

$$\Phi_{BIC}(\theta, T) = \frac{|\theta|_0 \log\, T}{2T} \tag{9}$$

In a statistically plausible interpretation of the world, Rissanen [16] obtains for regret the shortest possible description of the model using the universal distribution function of Kolmogorov, hence the name *Minimum Description Length*,

$$\Phi_{MDL}(\theta, T) = Coding\ Length_{Universal\ \ Distribution}(Model(\theta, T)) \tag{10}$$

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

In this paper we propose the following regret function

$$\Phi_{MDL-BSS}(\theta, T) = log_2(L) \tag{11}$$

Thus the optimization in (7) is carried out in two steps. First, for fixed L, the log likelihood is optimized over θ':

$$\begin{aligned} \hat{\theta'}_{MLE}(L) &= argmax_{\theta'} P(X|\theta', L) \\ MLV(L) &= P(X|\hat{\theta'}_{MLE}, L) \end{aligned} \tag{12}$$

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

$$\hat{L}_{MDL-BSS} = armin_L\, [-\log(MLV(L)) \;+\; \log_2(L)] \tag{13}$$

In the next subsection we present the computation of the Maximum Likelihood Value (MLV). Then, in the following subsection we argue the particular form (11) for $\Phi(\theta, T)$ inspired by the MDL interpretation. In same subsection we also discuss difficulties in a straightforward application of AIC or BIC criteria.

A. The Maximum Likelihood Value

The material from this subsection is presented in more detail in [2]. Results are summarized here for the benefit of the reader.

The constraint (H2) assumed in section I can be recast by introducing the selection variable $V(t)$: $V(t) = l$ iff $S_l(t) \neq 0$, and the amplitudes $G(t)$. Thus a slightly different parametrization of the model is obtained. The new set of parameters is now $\psi = (\psi', L)$ where

$$\psi' = \left(\{(G(t), V(t)) \ ; \ 1 \leq k \leq T\} \ , \ \sigma^2\right) \quad (14)$$

The signals in θ' are simply obtained through: $S_{V(t)}(t) = G(t)$, and $S_l(t) = 0$ for $l \neq V(t)$.

The likelihood (3) becomes:

$$\mathcal{L}(\psi) = \frac{1}{(2\pi)^{DT/2}\sigma^{DT}} exp\left(-\frac{1}{2\sigma^2}\sum_t \|\mathbf{X}(t) - G(t)A_{V(t)}\|^2\right) \quad (15)$$

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

$$\begin{aligned} log\mathcal{L}((V)_t, L) &= -\frac{DT}{2} log(2\pi) - DT\, log(\sigma) \quad (16) \\ &- \frac{1}{2\sigma^2}\sum_t \left[\|\mathbf{X}(t)\|^2 - |\langle \mathbf{X}(t), A_{V(t)}\rangle|^2\right] \end{aligned}$$

The optimization over $(V)_t$ and A is performed by an algorithm similar to the K-means algorithm:

- For a fixed matrix A the optimal selection variables are

$$V(t) = argmax_m |\langle \mathbf{X}(t), A_m\rangle| \quad (17)$$

- For a fixed selection map $(V(t))_{k,\omega}$, consider the induced partition $\Pi_m = \{t \ ; \ V(t) = m\}$. Then A_m is obtained as the principal eigenvector of the covariance matrix

$$R_m = \sum_{t \in \Pi_m} x(t)x(t)^* \quad (18)$$

Thus $R_m A_m = \lambda_{max} A_m$.

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

$$\hat{\sigma^2}_{MLE} = \frac{1}{T}\sum_{t=1}^{T}\left[\|\mathbf{X}(t)\|^2 - |\langle \mathbf{X}(t), \hat{A}_{\hat{V}_{MLE}(t)}\rangle|^2\right] \quad (19)$$

Finally, the log maximum likelihood value becomes:

$$\begin{aligned} log(MLV(L)) &= \frac{1}{T} log(\mathcal{L}(\hat{\psi'}_{MLE}; L)) \quad (20) \\ &= -\frac{D}{2} log(2\pi) - 1/2 - \frac{D}{2}\, log(\hat{\sigma^2}_{MLE}) \end{aligned}$$

where $\hat{\psi'}_{MLE}$ denoted the optimal parameter set ψ' containing the combined optimal values $(\hat{V}_{MLE}(t))_t$, $(\hat{G}_{MLE}(t))_t$, $(\hat{A}_l)_{1\leq l\leq L}$, $\hat{\sigma^2}_{MLE}$.

B. 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-H4). In general this is an impossible task since the Kolmogorov's complexity function is unkown. However the L-dependent part of the model description is embodied in the mixing parameters $(A_l)_{1\leq l\leq L}$, and the selection map $(V(t))_t$. First we model the mixing matrix by a uniform distribution in the space of unit vectors with a finite discretization of, say, M levels. Then we discretize the source amplitudes by Q levels, and we give no prior preferential treatment of one source signal versus the others. Thus 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:

$$\begin{aligned} l^*(Model; T) &\leq Llog_2(M) + T\,(log_2(L) + log_2(Q)) \quad (21) \\ &\quad + C(Model) \end{aligned}$$

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 T, and then, for a given (L, T) the entropic length of ψ. This clearly represents only one possible way of encoding the pair $(Model(\psi), T)$.

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

$$\begin{aligned}\Phi_{MDL-BSS}(L,T) &= \frac{L\,log_2(M)+T\,(log_2(L)+log_2(Q))}{T} \\ &= log_2(L)+\frac{L\,log_2(M)}{T}+log_2(Q)\end{aligned} \quad (22)$$

which, for large T reduces to (11).

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 $LT+L+1$, whereas in ψ description, this number is only $2T+L+1$. 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: an amplitude, and a selection map V. Consequently, the regret function for AIC can be either $L+\frac{L+1}{T}$, or $2+\frac{L+1}{T}$. Similarly, for BIC the regret function can be $L\,log(T)/2+\frac{(L+1)log(T)}{2T}$, or $log(T)+\frac{(L+1)log(T)}{2T}$. The criterion we propose in (22) interpolates between these two extrema, and captures better the actual size of the model parametrization.

III. Experimental Evaluation

We have generated mixing data according to model (1) with the following parameters. The number of sensors (D) varied from 2 to 4. The number of sources L ran from 1 to 8. Signal amplitude was generated as a unit variance Gaussian random variable. At every time t, the activation map was generated as a uniform random variable ranging from 1 to L. Gaussian noise was added to mixture with five levels of signal-to-noise ratio: 0, 25, 50, 75, and 100 dB.

There were 10 experiments, and in each experiment 10000 samples of signal were generated. Two signal traces for $D=2$, $L=3$, and two levels of SNR (0dB on top, and 100dB on the bottom) are rendered in Figure 1. Since the correction terms are of order $log(L+1)/T=10^{-4}$. the only meaningful AIC and BIC were given by the former regret functions. To summarize, the source number estimator is given by:

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

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

$$\hat{L}_{BIC} = argmin_L\,[-log\,MLV(L)+L\,log(T)/2] \quad (25)$$

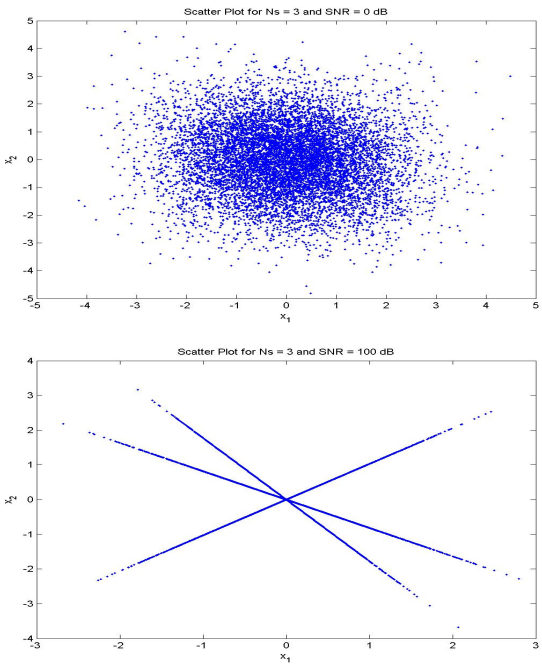


Fig. 1. Two scatter plots for $D=2$, $L=3$ and $SNR=0$dB (top), respectively $SNR=100$dB (bottom).

For an experiment with $D=2$ sensors, $L=5$ sources, and $SNR=100dB$, the plot of Maximum Log Likelihood, Regret, and combined criterion is plot in Figure 2; top plot refers to the MDL criterion, middle plot refers to the AIC criterion, bottom plot relates to the BIC criterion.

For a total of 1200 experiments (5 levels of noise x 3 choices of array size x 8 number of sources x 10 realizations), the histogram of estimation error has been obtained. For each of the three estimators, the histogram is rendered in Figure 3. Statistical performance of these estimators is presented in Table at right.

IV. 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.

This paper provides a theoretical framework for a statistical criterion to estimate number of source signals in a linear instantaneous mixing scenario with sparse signals. The numerical simulations confirmed the estimation performance.

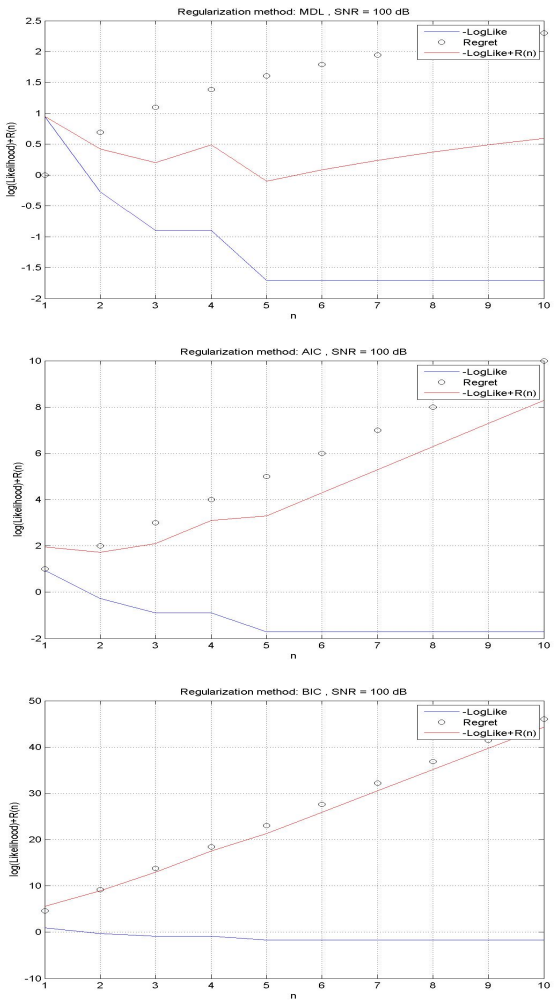
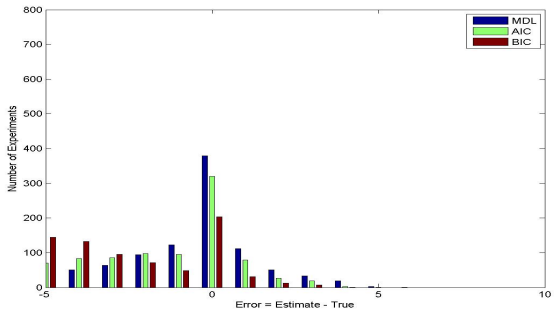


Fig. 2. MDL (top), AIC (middle), and BIC (bottom) related plots for $D = 2$, $L = 5$ and $SNR = 100$ dB.



Algorithm	Bias	Variance	Probability of Error
MDL-BSS	-0.693	4.71	62 %
AIC	-1.768	6.35	68 %
BIC	-3.254	6.82	77 %

Fig. 3. 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.

REFERENCES

[1] A. Belouchrani and M. Amin, "Blind source separation based on time-frequency signal representations," *IEEE Trans. on Signal Processing*, vol. 46, no. 11, pp. 2888–2897, Nov. 1998.

[2] S. Rickard, R. Balan, and J. Rosca, "Blind source separation based on space-time-frequency diversity," in *Proceedings of the 4th International Conference on Independent Component Analysis and Blind Source Separation (ICA2003), Nara, Japan*, April 2003.

[3] 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.

[4] A. Cichocki, Y. Li, P. Georgiev, and S.-I. Amari, "Beyond ica: Robust sparse signal representations," in *IEEE ISCAS Proc.*, 2004, pp. 684–687.

[5] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, Wiley, April 2002.

[6] R. Balan, "Estimator for number of sources using minimum description length criterion for blind sparse source mixtures," in *Proc. BSS-ICA*, 2007.

[7] S. Rickard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation," in *Proc. ICA*, 2001, pp. 651–656.

[8] 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.

[9] 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.

[10] J. Rosca, C. Borss, and R. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *Proc. ICASSP*, 2004.

[11] R. Balan and J. Rosca, "Convolutive demixing with sparse discrete prior models for markov sources," in *Proc. BSS-ICA*, 2006.

[12] R. Balan and J. Rosca, "Map source separation using belief propagation networks," in *Proc. ASILOMAR*, 2006.

[13] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Aut. Cont.*, vol. 19, no. 6, pp. 716–723, 1974.

[14] 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.

[15] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[16] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.