ORDER DETECTION FOR DEPENDENT SAMPLES USING ENTROPY RATE

Gengshen Fu, Hualiang Li, Matthew Anderson and Tülay Adalı

University of Maryland, Baltimore County, Dept. of CSEE, Baltimore, MD 21250

ABSTRACT

Detecting the number of signals in a given number of observations, or order detection, is one of the key issues in many signal processing problems. Information theoretic criteria are widely used to estimate the order. In many applications, data does not follow the independently and identically distributed (i.i.d.) sampling assumption. Previous approaches address dependent samples by downsampling the dataset so that existing order detection methods can be used. By downsampling the data, the sample size is decreased so that the accuracy of the order estimation is degraded. In this paper, we introduce two linear mixture models with dependent samples. The likelihood for each model is developed based on the entire data set and used in an information theoretic framework to improve the order estimation performance for dependent samples. Experimental results show performance improvement using this new method.

Index Terms— Order detection, Entropy rate, MDL criteria.

1. INTRODUCTION

A key issue when using a linear signal mixture with additive noise model is the detection of the number of signals. The seminal paper [1] introduced the use of information theoretic criteria (ITC) for the order selection problem. The ITC methods of [1] are now widely used for order selection because the order is estimated without requiring the user to specify a subjective threshold. In [1], the samples are assumed to be i.i.d. in order to calculate the likelihood. For many applications, the i.i.d. sampling assumption is violated. This paper focuses on order estimation without an i.i.d. sampling assumption, i.e., when the samples are dependent.

A reasonable assumption in practice is that the sample dependence is finite, i.e., there is finite memory in the observations. In the sequel, we restrict ourselves to finite memory, which allows for dependency to be removed by downsampling as done in previous approaches that address the problem. An intuitive method using an entropy rate matching principle is proposed in [2] to estimate the downsampling depth and then estimate the number of signals using [1]. There are several drawbacks for this method. First, a subjective threshold is needed to estimate the downsampling depth. Second, this method only uses a subset of the available samples to estimate the order. For example, only 10% of data will be used if downsampling depth is 10. Third, the calculated likelihood is for the downsampled data set, not for the whole data set. An improved method is given in [3]. The downsampling depth is estimated by using ITC to avoid specifying any subjective threshold. As in [2], the downsampled data is used to calculate the likelihood. Since the maximum likelihood estimator has large sample optimality properties and can be heavily biased when using small sample sizes, using all the samples in the estimation is likely to significantly improve the performance as we show here.

In this paper, we introduce two different models for generating sample dependence. Then we calculate their respective likelihoods, using *all* the available samples, and determine how many free parameters are available, as required to implement ITC methods. Also, we show that this new method is a generalization of the method given in [1]. Simulation results show this new method can improve the performance, especially for small data sets, low signal to noise ratios (SNR), and high memory lengths.

2. BACKGROUND

2.1. Linear Mixture Model

We consider the commonly assumed linear signal model with additive noise:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t), \quad t = 1, \ldots, T, \tag{1}$$

where t is the discrete time index, $\mathbf{x}(t)$ is the $N \times 1$ complex-valued observed random vector of the tth sample, $\mathbf{A}$ is a full column rank $N \times M$ complex-valued mixing matrix with $0 \leq M < N$, $\mathbf{s}(t)$ is the $M \times 1$ complex-valued latent Gaussian source vector, $\mathbf{n}(t)$ is an $N \times 1$ complex-valued white Gaussian noise vector, which is isotropic. In this paper, signal and noise are only considered to be circular complex, so that the *complete* second-order statistics are captured by the ordinary covariance matrices. Sample size is T. The problem is to detect model order M. The compact form for (1) is $\mathbf{X} = \mathbf{AS} + \mathbf{N}$, where $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(T)]$, $\mathbf{S} = [\mathbf{s}(1), \ldots, \mathbf{s}(T)]$ and $\mathbf{N} = [\mathbf{n}(1), \ldots, \mathbf{n}(T)]$.

2.2. ITC Based Order Estimation

Order estimation via ITC can be achieved using, among others, Akaike's Information Criterion (AIC) [4], Bayesian In-

This work was supported by the NSF grants NSF-CIF 1117056.

formation Criterion (BIC) [5] or equivalently Minimum Description Length (MDL) [6]. A general form for ITC is given by

$$\hat{M} = \arg\min_{M}\{-\alpha \log P(\mathbf{x}(1), \ldots, \mathbf{x}(T)|\boldsymbol{\theta}_M) + r(\boldsymbol{\theta}_M)\eta(T)\},$$

where $P(\mathbf{x}(1), \ldots, \mathbf{x}(T)|\boldsymbol{\theta}_M)$ is the joint probability density function, $r(\boldsymbol{\theta}_M)$ indicates all free parameters for order M. For AIC, $\alpha = 2$ and $\eta(T) = 2$. For MDL and BIC, $\alpha = 1$ and $\eta(T) = 0.5 \log T$. For model (1) with i.i.d. samples, it is shown in [1] that order detection using MDL is consistent as $T \to \infty$, and the AIC is shown to be inconsistent. Hence, we use the MDL criterion in this paper. If $\mathbf{x}$ is i.i.d., Gaussian distributed, then the MDL order selection criterion [1, 7] is

$$J_{\lambda} = \sum_{i=1}^{M} \log \lambda_i^2 + (N-M)\log \sigma^2 + \frac{M(2N-M)\log T}{2T}, \quad (2)$$

where λ_i is the ith largest eigenvalue of $\hat{\mathbf{C}}_x = \mathbf{X}\mathbf{X}^H/T$, $\sigma^2 = \frac{1}{N-M}\sum_{i=M+1}^{N} \lambda_i^2$, and superscript ^{H} denotes conjugate transpose.

3. LOG LIKELIHOOD CALCULATION FOR DEPENDENT SAMPLES

The unitary linear transform $\mathbf{Y} = \mathbf{U}^H\mathbf{X}$ yields (second-order) uncorrelated principal components $[\mathbf{y}_1, \ldots, \mathbf{y}_N]$, where $\mathbf{U}$ is the eigenvector matrix of $\hat{\mathbf{C}}_x$. This holds when the sample size $T \to \infty$, even for dependent samples. Assuming $\mathbf{y}_i$ is stationary and has finite memory of length K_i, the likelihood function is given by

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= -\frac{1}{T}\log P(\mathbf{X}) && (3)\\
&= \sum_{i=1}^{N}\Big(-\frac{1}{T}\log P(\mathbf{y}_i)\Big) && (4)\\
&= \sum_{i=1}^{N}\Big(-\frac{1}{T}\log\{P[y_i(1)]\cdot P[y_i(2)|y_i(1)]\cdot\cdots\\
&\qquad \cdot P[y_i(T)|y_i(T-1), \ldots, y_i(T-K_i+1)]\}\Big) && (5)\\
&= \sum_{i=1}^{N}\frac{-1}{T}\Big(\sum_{t=1}^{T_i'}\log P[\mathbf{y}_i(t,K_i)] - \sum_{t=2}^{T_i'}\log P[\mathbf{y}_i(t,K_i-1)]\Big)\\
&\approx \sum_{i=1}^{N}\frac{-1}{T_i'}\Big(\sum_{t=1}^{T_i'}\log P[\mathbf{y}_i(t,K_i)] - \sum_{t=1}^{T_i'}\log P[\mathbf{y}_i(t,K_i-1)]\Big)\\
& && (6)\\
&= \frac{N}{2}\log(2\pi) + \frac{1}{2}\sum_{i=1}^{N}\Big(\Big(\log\det(\mathbf{C}_{i(K_i)}) + \mathrm{tr}(\mathbf{C}_{i(K_i)}^{-1}\hat{\mathbf{C}}_{i(K_i)})\Big)\\
&\qquad -\Big(\log\det(\mathbf{C}_{i(K_i-1)}) + \mathrm{tr}(\mathbf{C}_{i(K_i-1)}^{-1}\hat{\mathbf{C}}_{i(K_i-1)})\Big)\Big), && (7)
\end{aligned}$$

where $T_i' \triangleq T - K_i + 1$, and $\mathbf{y}_i(t, L) \triangleq [y_i(t), \ldots, y_i(t+L-1)]^T$. The $K_i \times K_i$ and $(K_i-1)\times(K_i-1)$ autocorrelation matrix of $\mathbf{y}_i$ are given by $\mathbf{C}_{i(K_i)}$ and $\mathbf{C}_{i(K_i-1)}$. The corresponding sample autocorrelation matrices $\hat{\mathbf{C}}_{i(K_i)}$ and $\hat{\mathbf{C}}_{i(K_i-1)}$ are defined by

$$\hat{\mathbf{C}}_{i(L)} \triangleq \frac{1}{T_i'}\sum_{t=1}^{T_i'}\mathbf{y}_i(t,L)\mathbf{y}_i^H(t,L),$$

where $L = K_i$ and $K_i - 1$ respectively. Since $\mathbf{y}_i$ is Gaussian distributed, uncorrelatedness implies independence. Hence, (4) follows (3). By the assumption that $\mathbf{y}_i$ has finite memory K_i, we have (5). When the sample size $T \to \infty$, the ratio $(T/T') \to 1$, thus the approximation in (6) holds. By the assumption of stationarity of $\mathbf{y}_i$, we arrive at (7) from (6). Using $\hat{\mathbf{C}}_{i(K_i)}$ and $\hat{\mathbf{C}}_{i(K_i-1)}$ as the estimation of $\mathbf{C}_{i(K_i)}$ and $\mathbf{C}_{i(K_i-1)}$ and omitting terms that do not depend on the parameters, we can write the likelihood function $\mathcal{L}$ as

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N}\frac{1}{2}\log\frac{\det(\hat{\mathbf{C}}_{i(K_i)})}{\det(\hat{\mathbf{C}}_{i(K_i-1)})}. \quad (8)$$

We can also understand the previous derivation from the entropy rate perspective. Under the assumption of stationary ergodic process y, from the general asymptotic equipartition property, we know

$$-\log P(y(1), \ldots, y(T)) \to H(y(1), \ldots, y(T)),$$

and the definition of entropy rate is

$$\mathcal{H}(\mathbf{y}) \triangleq \lim_{T\to\infty}\frac{H(y(1), \ldots, y(T))}{T}.$$

Then $\frac{1}{2}\log(2\pi e)\frac{\det(\mathbf{C}_{i(K_i)})}{\det(\mathbf{C}_{i(K_i-1)})} = -\frac{1}{T}\log P(\mathbf{y}_i)$ can be interpreted as the entropy rate of $\mathbf{y}_i$. That is to say, the estimation of the likelihood function is equivalent to the entropy rate estimation. Let $h_i \triangleq \det(\mathbf{C}_{i(K_i)})/\det(\mathbf{C}_{i(K_i-1)})$, which is a one-to-one monotonic mapping of the entropy rate of $\mathbf{y}_i$.

4. MEMORY LENGTH ESTIMATION

From the previous discussion, we know that memory lengths, $K_i, i = 1, \ldots, N$, are needed to calculate the likelihood. Given a sequence $\{z(t)\}, t = 1, \ldots, T$, the problem of memory length estimation is the same as finding a minimum downsampling rate K, such that the downsampled sequence $\{z_K(t)\}$ is i.i.d.. A hypotheses test method is proposed by [3] to estimate the minimum downsampling rate for real-valued data. We extend this method to complex-valued data. By assuming $\{z_K(t)\}$ is generated by an autoregressive (AR) model, we have the following hypothesis test:

H_0 : The AR order of $\{\bar{z}_{d,r}(n)\}$ is zero.
H_1 : The AR order of $\{\bar{z}_{d,r}(n)\}$ is positive.

Hypothesis H_0 means $\{z_K(t)\}$ is i.i.d., otherwise the samples are dependent. We use the MDL criterion to estimate the AR order q as [8]:

$$\hat{q} = \arg\min_{q}\left\{T_d \log \sigma_q^2 + q\log(T_d)\right\},$$

where σ_q^2 is the variance of the prediction error, and T_d is the downsampled sample size.

5. ORDER ESTIMATION

From the point of view of principal component analysis, the signal subspace should be the mixture of signal and noise, and the noise subspace should only contain noise. So, for the i.i.d. case, the signal subspace has higher energy than the noise subspace. For the case with dependent samples, this property translates to the signal subspace having higher entropy rate than the noise subspace. We thus distinguish the two subspaces via this property. Without loss of generality, we assume $\mathbf{y}_i$ is zero mean. To address sample dependence in the observations, we consider the following two models.

5.1. Model by entropy rate

For the first dependency model the entropy rates are assumed to be the same in the noise subspace.

$$\begin{array}{ll} \mathbf{y}_1 \sim \mathcal{N}:\{0, h_1\} & \mathbf{y}_{M+1} \sim \mathcal{N}:\{0, \bar{h}\} \\ \dots & \dots \\ \mathbf{y}_M \sim \mathcal{N}:\{0, h_M\} & \mathbf{y}_N \sim \mathcal{N}:\{0, \bar{h}\} \end{array}$$
$$\mathbf{X} = \mathbf{U}\mathbf{Y},$$

where $\mathcal{N}:\{0,h\}$ refers to zero mean Gaussian random process with entropy rate h. The parameters for this model are $\{h_1,\dots,h_M,\bar{h},\mathbf{U}\}$. Using the likelihood in (8), the MDL criterion for this model is

$$J_{\mathrm{h}} = \sum_{i=1}^{M} \log h_i + (N-M)\log \bar{h} + \frac{M(2N-M)\log T}{2T} \quad (9)$$

and the estimators for h_i and $\bar{h}$ are given by $\hat{h}_i = \frac{\det(\hat{\mathbf{C}}_{i(K_i)})}{\det(\hat{\mathbf{C}}_{i(K_i-1)})}$ and $\hat{\bar{h}} = \frac{1}{N-M}\sum_{i=M+1}^{N}\hat{h}_i$ respectively.

5.2. Model by autocorrelation matrix

For the second dependency model, the autocorrelation matrices are assumed to be the same in the noise subspace.

$$\begin{array}{ll} \mathbf{y}_1 \sim \mathcal{N}:\{0, \mathbf{C}_1\} & \mathbf{y}_{M+1} \sim \mathcal{N}:\{0, \bar{\mathbf{C}}\} \\ \dots & \dots \\ \mathbf{y}_M \sim \mathcal{N}:\{0, \mathbf{C}_M\} & \mathbf{y}_N \sim \mathcal{N}:\{0, \bar{\mathbf{C}}\} \end{array}$$
$$\mathbf{X} = \mathbf{U}\mathbf{Y},$$

where $\mathcal{N}:\{0,\mathbf{C}_i\}$ refers to zero mean Gaussian random process with a $K_i \times K_i$ autocorrelation matrix $\mathbf{C}_i$. The parameters for this model are $\{\mathbf{C}_1,\dots,\mathbf{C}_M,\bar{\mathbf{C}},\mathbf{U}\}$. Since $\mathbf{C}_i$ has Toeplitz structure and is complex, it has $2K_i - 1$ free parameters. Using the likelihood in (8), the MDL criterion for this model is

$$J_{\mathrm{C}} = \sum_{i=1}^{M} \log \frac{\det(\mathbf{C}_{i(K_i)})}{\det(\mathbf{C}_{i(K_i-1)})} + (N-M)\log\frac{\det(\bar{\mathbf{C}}_{(\bar{K})})}{\det(\bar{\mathbf{C}}_{(\bar{K}-1)})} + \frac{\left(M(2N-M-2) + \sum_{i=1}^{M} 2K_i + 2\bar{K}\right)\log T}{2T}, \quad (10)$$

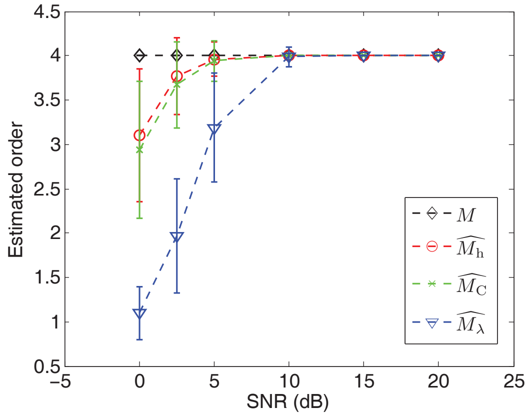

Fig. 1. Performance comparison as a function of SNR.

and the estimation for $\hat{\bar{\mathbf{C}}}$ is $\hat{\bar{\mathbf{C}}} = \frac{1}{N-M}\sum_{i=M+1}^{N}\hat{\mathbf{C}}_i$.

Both sample dependency models are generalizations of the i.i.d. model. Order is estimated using a likelihood that is formed using the whole data set. If samples are i.i.d., $\bar{K} = 1, K_i = 1, i = 1\dots M$, then $h_i = \mathbf{C}_i = \lambda_i^2$ and $\bar{h} = \bar{\mathbf{C}} = \sigma^2$. For this case, both (9) and (10) reduce to (2). Comparing with the model by entropy rate, model by autocorrelation matrix is more specific. The assumption that autocorrelation matrices are same in noise subspace requires not only entropy rate but also dependency structure to be same in the noise subspace, and more parameters need to be estimated.

6. EXPERIMENTAL RESULTS

In this section, we use simulations to study the properties of the order detection method we introduced that uses entropy rate. We use a moving average model to simulate the observed samples. The experiment default settings are $M = 4$ Gaussian signal vectors, $N = 10$ Gaussian noise vectors, memory lengths $K_i = 10, i = 1,\dots,M, \bar{K} = 10$, sample size $T = 1000$, and SNR $= 5$ dB. For each experiment below we vary one parameter of the default settings. Then we compare the order estimation performance of

- $\widehat{M}_{\mathrm{h}}$: Order estimated by (9).
- $\widehat{M}_{\mathrm{C}}$: Order estimated by (10).
- $\widehat{M}_{\lambda}$: Order estimated by the downsampling method proposed in [3], as this method is shown to perform better than both the methods given in [2] and [1] when samples are not i.i.d.

Experiment 1: In this experiment, we compare the performance as we vary the SNR. From Fig. 1, we can see the performance of all three methods degrade as SNR decreases, and $\widehat{M}_{\mathrm{h}}$ yields the best performance.

Experiment 2: In this experiment, we compare the performance as we vary the sample size T. As shown in Fig. 2, we can see, for all three methods, that larger sample size leads to better performance, and $\widehat{M}_{\mathrm{h}}$ yields the best performance.

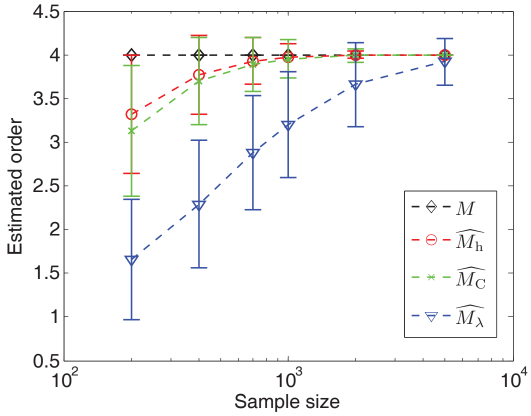

Fig. 2. Performance comparison as a function of sample size.

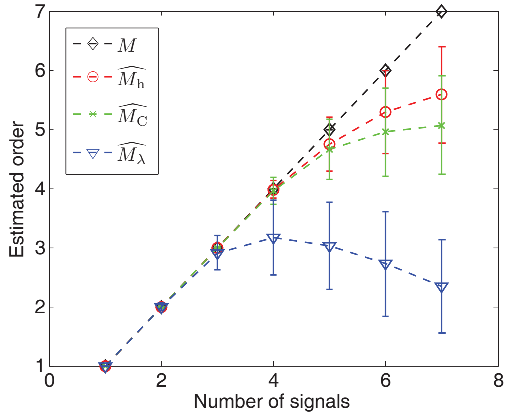

Fig. 3. Performance comparison as a function of number of signals.

Experiment 3: Next, we compare the performance as we vary the signal number M. As observed from Fig. 3, the performance of all three methods degrade with increasing signal number, and again $\widehat{M_{\mathrm{h}}}$ yields the best performance.

Experiment 4: We compare the performance as we vary the memory length K. From Fig. 4, as the memory length increases, the performance of all three methods deteriorate. However, the best performance still achieved by $\widehat{M_{\mathrm{h}}}$.

As observed in these four experiments, the performance of $\widehat{M_{\mathrm{h}}}$ is always better than the other two order estimators, especially when the sample size is small, the SNR is low, and the memory length is high. As proved in [1], the MDL criterion is consistent as the sample size $T \to \infty$. So, the performance of the method proposed in [3] is poor because the sample size used for order detection decreases significantly after downsampling. The reason that estimation performance of $\widehat{M_{\mathrm{h}}}$ is better than $\widehat{M_{\mathrm{C}}}$ is that estimation of h_i, which is a scalar, is more likely to be accurate than estimation of $\mathbf{C}_i$, which is a matrix and has more parameters need to be estimated.

7. DISCUSSION

In this paper, we propose a new order estimation method for dependent samples, and provide its interpretation using entropy rate. We show by experiment that $\widehat{M_{\mathrm{h}}}$ yields best performance among the two models proposed in this paper and the method proposed in [3], which has been shown to be superior to the previous approaches. Although we only consider circular-valued complex data in this paper, the new method can be extended to noncircular complex-valued data.

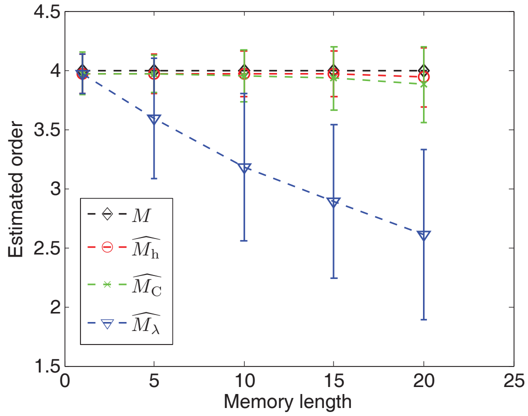

Fig. 4. Performance comparison as a function of memory length.

8. REFERENCES

[1] M. Wax and T. Kailath, “Detection of signals by information theoretic criteria,” *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 33, no. 2, pp. 387–392, 1985.

[2] Y. O. Li, T. Adalı, and V. D. Calhoun., “Estimating the number of independent components for fMRI data,” *Hum. Brain Mapp.*, vol. 28, no. 11, pp. 1251–1266, 2007.

[3] X. L. Li, S. Ma, V. D. Calhoun, and T. Adalı, “Order detection for fMRI analysis: Joint estimation of downsampling depth and order by information theoretic criteria,” *IEEE Int. Symp., Biomedical Imaging: From Nano to Macro*, pp. 1019–1022, 2011.

[4] H. Akaike, “Information theory and an extension of the maximum likelihood principle.” *in Proc. 2nd Int. Symp. on Information Theory, Tsahkadsor, Armenian SSR; Hungary*, pp. 267–281, 1973.

[5] G. Schwarz, “Estimating the dimension of a model.” *Ann. Stat.*, vol. 6, pp. 461–464, 1978.

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

[7] X. L. Li, T. Adalı, and M. Anderson, “Noncircular principal component analysis and its application to model selection,” *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4516–4528, oct. 2011.

[8] E. J. Hannan, “The estimation of the order of an ARMA process,” *Ann. Stat.*, vol. 8, no. 5, pp. 1071–1081, 1980.