Inconsistency of the MDL: On the Performance of Model Order Selection Criteria with Increasing Signal-to-Noise Ratio

Quan Ding, *Student Member, IEEE,* and Steven Kay, *Fellow, IEEE*

Abstract

In the problem of model order selection, it is well known that the widely used minimum description length (MDL) criterion is consistent as the sample size $N \to \infty$. But the consistency as the noise variance $\sigma^2 \to 0$ has not been studied. In this paper, we find that the MDL is inconsistent as $\sigma^2 \to 0$. The result shows that the MDL has a tendency to overestimate the model order. We also prove that another criterion, the exponentially embedded family (EEF), is consistent as $\sigma^2 \to 0$. Therefore in a high signal-to-noise (SNR) scenario, the EEF provides a better criterion to use for model order selection.

Index Terms

Consistency, exponentially embedded families, hypothesis testing, minimum description length, model order selection

I. Introduction

Model order selection is a fundamental problem in signal processing. It has many practical applications such as radar, computer vision and biomedical systems. Model order selection is essentially one of composite hypothesis testing, for which the probability density functions (PDFs) are known except for some parameters. Without the knowledge of those parameters, there exists no optimal solution. A simple and common approach is the generalized likelihood ratio test (GLRT) which replaces the unknown parameters by their maximum likelihood estimates (MLEs). However in the case when the

This work was supported by the Air Force Office of Scientific Research under Contract AFOSR FA9550-08-1-0459.

Quan Ding and Steven Kay are with the Department of Electrical Computer and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881, USA. E-mail: dingqqq@ele.uri.edu, kay@ele.uri.edu.

model orders are hierarchically nested, the GLRT philosophy does not work since it will always choose the largest candidate order (see [1] for a simple example). Many methods have been proposed to offset this overestimating tendency based on different information criteria such as the Akaike's information criterion (AIC) [2], the MDL [3], [4], and the EEF [5]. The reader may wish to read [6] for a review of information criterion rules on model order selection. One would prefer a criterion that will always choose the true model order if we have a large enough number of samples. It has been shown in [7] the consistency of the MDL and the inconsistency of the AIC as the sample size $N \to \infty$, i.e., the MDL will pick the true order with probability one and the AIC tends to overestimate the model order as $N \to \infty$. The consistency of the EEF as $N \to \infty$ is shown in [8].

Except for the above consistency as $N \to \infty$, one would also wish the criterion to have another consistency that we call *consistency as* $\sigma^2 \to 0$. In this case the estimator will choose the true model order in probability as the noise level decreases to zero. This is the consistency that we will discuss throughout this paper. The Fisher consistency [9] is the same as the consistency as $\sigma^2 \to 0$ in parameter estimation in curved exponential families [10]. To our knowledge, no work has been done on the consistency as $\sigma^2 \to 0$ for the model order selection criteria. In this paper, we will show that the MDL and the AIC are inconsistent as the noise variance $\sigma^2 \to 0$. This means that even under high SNR conditions, the MDL and the AIC still tend to overestimate the model order. We then show that the EEF is consistent as $\sigma^2 \to 0$. Simulation results are provided to support our analysis.

The paper is organized as follows. Section II presents the problem and the model order selection criterions. Then we introduce a linear model and show the inconsistency as $\sigma^2 \to 0$ for the MDL and the AIC in Section III. In Section IV, we prove that the EEF is consistent as $\sigma^2 \to 0$. Simulation results are given in Section V to justify our derivation. Finally, Section VI draws the conclusion.

II. Problem Statement

Consider the multiple composite hypothesis testing problem where we have M candidate models. Under each model $\mathcal{H}_i$, we have

$$\mathcal{H}_i : \mathbf{x} = \mathbf{s}_i(\boldsymbol{\theta}_i) + \mathbf{w} = \mathbf{s}_i(\boldsymbol{\theta}_i) + \sigma\mathbf{u} \tag{1}$$

for $i = 1, 2, \ldots, M$. $\mathbf{x}$ is an $N \times 1$ vector of samples. The $N \times 1$ signal $\mathbf{s}_i(\boldsymbol{\theta}_i)$ is known except for the unknown $i \times 1$ vector of parameters $\boldsymbol{\theta}_i$. $\mathbf{w} = \sigma\mathbf{u}$ is the $N \times 1$ noise vector with known variance σ^2, and $\mathbf{u}$ has a well defined PDF. So each $\mathcal{H}_i$ is described by a PDF $p(\mathbf{x}; \boldsymbol{\theta}_i)$. We assume that the model orders

are hierarchically nested, i.e., we can write the signal $\mathbf{s}_i(\boldsymbol{\theta}_i)$ as

$$\mathbf{s}_i(\boldsymbol{\theta}_i) = \mathbf{s}\left([\theta_1, \ldots, \theta_i, 0, \ldots]^T\right) \tag{2}$$

where $\mathbf{s}$ is a function of a $M \times 1$ vector, for $i = 1, 2, \ldots, M$. So the unknown parameters in signal with higher order contain all of those in a lower order model. Let $\mathcal{H}_0$ be a reference hypothesis with $\mathbf{s}\left([0, 0, \ldots, 0]^T\right) = \mathbf{0}$, so the PDF $p(\mathbf{x}; \boldsymbol{\theta}_0)$ is completely known as noise only. Then the MDL, AIC and EEF rules choose the model order that maximizes the following respectively:

$$\begin{aligned}
&- MDL(i) = l_{G_i}(\mathbf{x}) - i \ln N \\
&- AIC(i) = l_{G_i}(\mathbf{x}) - 2i \\
&EEF(i) = \left(l_{G_i}(\mathbf{x}) - i\left[\ln\left(\frac{l_{G_i}(\mathbf{x})}{i}\right) + 1\right]\right) u\left(\frac{l_{G_i}(\mathbf{x})}{i} - 1\right)
\end{aligned}$$

for $i = 1, 2, \ldots, M$, where $u(x)$ is the unit step function and $l_{G_i}(\mathbf{x}) = 2 \ln \frac{p(\mathbf{x}; \hat{\boldsymbol{\theta}}_i)}{p(\mathbf{x}; \boldsymbol{\theta}_0)}$. Here $\hat{\boldsymbol{\theta}}_i$ is the MLE for $\boldsymbol{\theta}_i$. Note that the inclusion of the term $-2 \ln p(\mathbf{x}; \boldsymbol{\theta}_0)$ does not affect the maximum and so we use the log-likelihood ratio instead of the more usual log-likelihood for the MDL and the AIC. In the next section we will implement these rules in the linear model to show the inconsistency of the MDL and the AIC as $\sigma^2 \to 0$.

III. Inconsistency of the MDL and the AIC

Without causing any confusion, we will use *consistency* instead of *consistency as* $\sigma^2 \to 0$ for the rest of the paper unless otherwise mentioned. In this section, we will limit the derivation to the MDL. We will start by introducing the linear model, from which we derive the performance of the MDL. Then the inconsistency of the MDL is readily seen. The inconsistency of the AIC follows directly from the analysis of the MDL.

A. The Linear Model

Consider the following linear model:

$$\mathcal{H}_i : \mathbf{x} = \mathbf{H}_i \boldsymbol{\theta}_i + \mathbf{w} \quad \text{for } i = 1, 2, \ldots, M$$

where M is the maximum order of all the candidate models, $\mathbf{H}_i = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_i]$ is an $N \times i$ (with $N > M$) known observation matrix with full column rank, $\boldsymbol{\theta}_i = [\theta_1, \theta_2, \ldots, \theta_i]^T$ is an $i \times 1$ unknown parameter vector of the amplitudes, and $\mathbf{w}$ is an $N \times 1$ white Gaussian noise vector with known variance σ^2. For the linear model, $l_{G_i}(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{P}_i \mathbf{x}}{\sigma^2}$, where $\mathbf{P}_i = \mathbf{H}_i (\mathbf{H}_i^T \mathbf{H}_i)^{-1} \mathbf{H}_i^T$ is the projection matrix that

projects $\mathbf{x}$ onto the subspace V_i generated by $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_i$ [11]. So the MDL rule chooses the model order that minimizes:

$$\mathrm{MDL}(i) = -\frac{\mathbf{x}^T \mathbf{P}_i \mathbf{x}}{\sigma^2} + i \ln N \quad \text{for } i = 1, 2, \ldots, M$$

Let $y_i = \frac{\mathbf{x}^T \mathbf{P}_{i+1} \mathbf{x}}{\sigma^2} - \frac{\mathbf{x}^T \mathbf{P}_i \mathbf{x}}{\sigma^2}$ for $i = 1, 2, \ldots, M-1$ and we have the following theorem. (See Appendix A for the proof of Theorem 1)

Theorem 1 (PDF of y_j for $j \geq p$). *If the true model order is $\mathcal{H}_p$ ($p \leq M$), that is, $\theta_i = 0$ for all $i > k$, then the y_j's for all $j \geq p$ do not depend on $\boldsymbol{\theta}_p$ or σ^2, and they are independent and identically distributed (IID), each with a chi-square distribution with 1 degree of freedom.*

As we will show next, this theorem gives us a way to find a lower bound of the probability that the MDL will choose the wrong model order.

B. Inconsistency of the MDL

We will show that the probability of overestimation does not converge to zero as $\sigma^2 \to 0$.
If $\mathcal{H}_p$ ($p < M$) is true, then the probability that the MDL will choose the wrong model order is

$$\begin{aligned} P_e &= \Pr\{\mathcal{H}_j, j \neq p | \mathcal{H}_p\} \\ &= 1 - \Pr\{MDL(p) < MDL(j) \text{ for all } j \neq p | \mathcal{H}_p\} \\ &\geq 1 - \Pr\{MDL(p) < MDL(j) \text{ for all } j > p | \mathcal{H}_p\} \\ &= \Pr\{MDL(p) \geq MDL(j) \text{ for some } j > p | \mathcal{H}_p\} \end{aligned} \tag{3}$$

Since $MDL(j) - MDL(j+1) = y_j - \ln N$, for $j > p$, $MDL(p) - MDL(j) = \sum_{i=p}^{j-1} y_i - (j-p) \ln N$, we have

$$\begin{aligned} &\Pr\{MDL(p) \geq MDL(j) \text{ for some } j > p | \mathcal{H}_p\} \\ &= \Pr\{y_p \geq \ln N \text{ or } y_p + y_{p+1} \geq 2 \ln N \text{ or } \cdots \text{ or } \sum_{i=p}^{M-1} y_i \geq (M-p) \ln N | \mathcal{H}_p\} \end{aligned} \tag{4}$$

By Theorem 1, $y_j \sim \chi_1^2$ and y_j's are independent for $j \geq p$. So the probability in (4) can be found analytically, although it may not easy. Alternatively, we can find a lower bound of (4) which is much easier to calculate. Notice that

$$\begin{aligned} &\Pr\{y_p \geq \ln N \text{ or } y_p + y_{p+1} \geq 2 \ln N \text{ or } \cdots \text{ or } \sum_{i=p}^{M-1} y_i \geq (M-p) \ln N | \mathcal{H}_p\} \\ &\geq \Pr\{y_p \geq \ln N | H_p\} = 2Q\left(\sqrt{\ln N}\right) \end{aligned} \tag{5}$$

where $Q(x)$ function is the right-tail probability of a standard Gaussian distribution, that is, $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right)dt$. So $2Q\left(\sqrt{\ln N}\right)$ is also a lower bound of the probability of error P_e for the MDL. Note that this lower bound decreases slowly as N increases. For example, in order to have $P_e \leq 0.01$, we require that $2Q\left(\sqrt{\ln N}\right) \leq 0.01$ and we need as many as $N = 761$ samples. This lower bound only depends on the number of samples N. So when N is fixed, this lower bound is fixed even as $\sigma^2 \to 0$. This shows that the MDL is inconsistent and it has a tendency to overestimate the model order.

For the AIC, we just need to replace $\ln N$ by 2, so the lower bound is $2Q\left(\sqrt{2}\right)$. Hence the AIC is also inconsistent. Notice that $2Q\left(\sqrt{\ln N}\right) \to 0$ as $N \to \infty$, but $2Q\left(\sqrt{2}\right)$ is a constant. This also justifies the result in [7]. Since the MDL is consistent as $N \to \infty$, the lower bound $2Q\left(\sqrt{\ln N}\right)$ should decrease to 0. The lower bound $2Q\left(\sqrt{2}\right)$ for the AIC shows that the AIC is inconsistent as $N \to \infty$.

IV. Consistency of the EEF

As a complement to Section III, we will first show that the EEF is consistent for the linear model. Next, we will prove that the EEF is consistent in general.

A. Consistency of the EEF for the Linear Model

The next theorem will be used to prove the consistency of the EEF for the linear model. (See Appendix B for the proof of Theorem 2)

Theorem 2 (PDF of y_j for $j < p$). *If the true model order is $\mathcal{H}_p$, then for $j < p$, y_j has a noncentral chi-square distribution with 1 degree of freedom and noncentrality parameter*

$$\lambda_j = \left(\mathbf{H}_{j+1,p}\boldsymbol{\theta}_{j+1,p}\right)^T \left(\mathbf{P}_{j+1} - \mathbf{P}_j\right) \mathbf{H}_{j+1,p}\boldsymbol{\theta}_{j+1,p}/\sigma^2$$

where $\mathbf{H}_{j+1,p} = [\mathbf{h}_{j+1}, \mathbf{h}_{j+2}, \ldots, \mathbf{h}_p]$ and $\boldsymbol{\theta}_{j+1,p} = [\theta_{j+1}, \theta_{j+2}, \ldots, \theta_p]^T$. Let

$$\alpha_j = \left(\mathbf{H}_{j+1,p}\boldsymbol{\theta}_{j+1,p}\right)^T \left(\mathbf{P}_{j+1} - \mathbf{P}_j\right) \mathbf{H}_{j+1,p}\boldsymbol{\theta}_{j+1,p}$$

and we have $\lambda_j = \alpha_j/\sigma^2$. Furthermore, the y_j's are independent for all j.

The EEF chooses the model order that maximizes

$$\begin{aligned} EEF(i) &= \left(l_{G_i}(\mathbf{x}) - i\left[\ln\left(\frac{l_{G_i}(\mathbf{x})}{i}\right) + 1\right]\right) u\left(\frac{l_{G_i}(\mathbf{x})}{i} - 1\right) \\ &= \left(\frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x}}{\sigma^2} - i\left[\ln\left(\frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x}}{i\sigma^2}\right) + 1\right]\right) u\left(\frac{\mathbf{x}^T\mathbf{P}_i\mathbf{x}}{i\sigma^2} - 1\right) \end{aligned} \tag{6}$$

If $\mathcal{H}_p$ is true, it is well known that [1]

$$l_{G_p}(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{P}_p \mathbf{x}}{\sigma^2} \sim \chi_p^{'2}(\lambda) \tag{7}$$

where $\lambda = \frac{\|\mathbf{H}_p \boldsymbol{\theta}_p\|^2}{\sigma^2}$. In order to prove the consistency of the EEF in probability, we need to show that

$$\Pr\left\{\arg\max_i EEF(i) = p\right\} \to 1$$

as $\sigma^2 \to 0$. We start by first comparing $EEF(j)$ with $EEF(p)$ as $\sigma^2 \to 0$ for $j > p$ and $j < p$.

For $j > p$, we know that [1]

$$l_{G_j}(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{P}_j \mathbf{x}}{\sigma^2} \sim \chi_j^{'2}(\lambda)$$

where λ is the same as in (7). The lemma in [8] shows that if $\mathbf{Y}$ is distributed according to $\chi_\nu^{'2}(an)$ where a is a positive constant, then as $n \to \infty$, $\frac{\mathbf{Y}}{n}$ converges to a in probability, or in symbols, $\frac{\mathbf{Y}}{n} \xrightarrow{P} a$. Replacing n by $1/\sigma^2$ we have as $\sigma^2 \to 0$

$$\begin{aligned} \sigma^2 l_{G_p}(\mathbf{x}) &\xrightarrow{P} \|\mathbf{H}_p \boldsymbol{\theta}_p\|^2 \\ \sigma^2 l_{G_j}(\mathbf{x}) &\xrightarrow{P} \|\mathbf{H}_p \boldsymbol{\theta}_p\|^2 \quad \text{for } j > p \end{aligned} \tag{8}$$

Therefore

$$\begin{aligned} \Pr\left\{\frac{l_{G_p}(\mathbf{x})}{p} - 1 > 0\right\} &\to 1 \\ \Pr\left\{\frac{l_{G_j}(\mathbf{x})}{j} - 1 > 0\right\} &\to 1 \quad \text{for } j > p \end{aligned} \tag{9}$$

as $\sigma^2 \to 0$ and we can discard the unit step function. As a result,

$$\begin{aligned} & EEF(k) - EEF(j) \\ &= l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x}) - p \ln l_{G_p}(\mathbf{x}) + j \ln l_{G_j}(\mathbf{x}) + p \ln p - j \ln j - p + j \\ &= l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x}) - p \ln \left(\sigma^2 l_{G_p}(\mathbf{x})\right) + j \ln \left(\sigma^2 l_{G_j}(\mathbf{x})\right) + c \end{aligned} \tag{10}$$

where

$$c = (p - j) \ln \sigma^2 + p \ln p - j \ln j - p + j \tag{11}$$

By Theorem 1,

$$l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x}) \sim -\chi^2_{j-p} \tag{12}$$

Since $\sigma^2 l_{G_p}(\mathbf{x}) \xrightarrow{P} \|\mathbf{H}_p \boldsymbol{\theta}_p\|^2, \sigma^2 l_{G_j}(\mathbf{x}) \xrightarrow{P} \|\mathbf{H}_p \boldsymbol{\theta}_p\|^2$, by the continuity of the logarithm we have [12]

$$\begin{aligned} \ln \left(\sigma^2 l_{G_p}(\mathbf{x})\right) &\xrightarrow{P} \ln \|\mathbf{H}_p \boldsymbol{\theta}_p\|^2 \\ \ln \left(\sigma^2 l_{G_j}(\mathbf{x})\right) &\xrightarrow{P} \ln \|\mathbf{H}_p \boldsymbol{\theta}_p\|^2 \quad \text{for } j > p \end{aligned} \tag{13}$$

We divide (10) by c and get

$$\frac{EEF(p)-EEF(j)}{c}=\frac{l_{G_p}(\mathbf{x})-l_{G_j}(\mathbf{x})-p\ln\left(\sigma^2 l_{G_p}(\mathbf{x})\right)+j\ln\left(\sigma^2 l_{G_j}(\mathbf{x})\right)}{c}+1$$

Since $\frac{1}{c}\to 0^+$ for $j>p$ as $\sigma^2\to 0$, as a result of (12) and (13), we have [12]

$$\frac{l_{G_p}(\mathbf{x})-l_{G_j}(\mathbf{x})-p\ln\left(\sigma^2 l_{G_p}(\mathbf{x})\right)+j\ln\left(\sigma^2 l_{G_j}(\mathbf{x})\right)}{c}\xrightarrow{P}0 \tag{14}$$

and hence

$$\frac{EEF(p)-EEF(j)}{c}\xrightarrow{P}1 \tag{15}$$

for $j>p$. This shows that as $\sigma^2\to 0$, $\Pr\{EEF(p)>EEF(j)\}\to 1$.

For $j<p$, similar to the derivation in Appendix B, the distribution of $l_{G_j}(\mathbf{x})=\frac{\mathbf{x}^T\mathbf{P}_j\mathbf{x}}{\sigma^2}$ can be found as

$$l_{G_j}\sim\chi_j^{'2}(\lambda') \tag{16}$$

where $\lambda'=\frac{\left(\mathbf{H}_p\boldsymbol{\theta}_p\right)^T\mathbf{P}_j\mathbf{H}_p\boldsymbol{\theta}_p}{\sigma^2}$. So we also have

$$\begin{aligned}&\Pr\left\{\frac{l_{G_p}(\mathbf{x})}{p}-1>0\right\}\to 1\\&\Pr\left\{\frac{l_{G_j}(\mathbf{x})}{j}-1>0\right\}\to 1\quad\text{for } j<p\end{aligned} \tag{17}$$

as $\sigma^2\to 0$. Thus we can also omit the unit step function and have

$$\begin{aligned}&EEF(p)-EEF(j)\\&=l_{G_p}(\mathbf{x})-l_{G_j}(\mathbf{x})-p\ln\left(\sigma^2 l_{G_p}(\mathbf{x})\right)+j\ln\left(\sigma^2 l_{G_j}(\mathbf{x})\right)+c\end{aligned} \tag{18}$$

where

$$c=(p-j)\ln\sigma^2+p\ln p-j\ln j-p+j \tag{19}$$

Now by Theorem 2,

$$l_{G_p}(\mathbf{x})-l_{G_j}(\mathbf{x})\sim\chi_{p-j}^{'2}\left(\sum\nolimits_{i=j}^{p-1}\lambda_i\right)=\chi_{p-j}^{'2}\left(\sum_{i=j}^{p-1}\alpha_i/\sigma^2\right) \tag{20}$$

so that by the lemma in [8], we have

$$\sigma^2\left(l_{G_p}(\mathbf{x})-l_{G_j}(\mathbf{x})\right)\xrightarrow{P}\sum_{i=j}^{p-1}\alpha_i \tag{21}$$

Similarly to the above analysis, we have

$$\begin{aligned}&\ln\left(\sigma^2 l_{G_p}(\mathbf{x})\right)\xrightarrow{P}\ln\|\mathbf{H}_p\boldsymbol{\theta}_p\|^2\\&\ln\left(\sigma^2 l_{G_j}(\mathbf{x})\right)\xrightarrow{P}\ln\left(\left(\mathbf{H}_p\boldsymbol{\theta}_p\right)^T\mathbf{P}_j\mathbf{H}_p\boldsymbol{\theta}_p\right)\quad\text{for } j<p\end{aligned} \tag{22}$$

Hence, with $\sigma^2 \to 0$ we have

$$\begin{aligned} &\sigma^2 \ln\left(\sigma^2 l_{G_p}(\mathbf{x})\right) \xrightarrow{P} 0 \\ &\sigma^2 \ln\left(\sigma^2 l_{G_j}(\mathbf{x})\right) \xrightarrow{P} 0 \quad \text{for } j < p \end{aligned} \tag{23}$$

Obviously, $\sigma^2 c \to 0$. So by (18), (21) and (23), we have

$$\begin{aligned} &\sigma^2 \left(EEF(p) - EEF(j)\right) \\ &= \sigma^2 \left(l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x})\right) - p\sigma^2 \ln\left(\sigma^2 l_{G_p}(\mathbf{x})\right) + j\sigma^2 \ln\left(\sigma^2 l_{G_j}(\mathbf{x})\right) + \sigma^2 c \\ &\xrightarrow{P} \sum_{i=j}^{p-1} \alpha_i > 0 \end{aligned} \tag{24}$$

for $j < p$. This means that $\Pr\{EEF(p) > EEF(j)\} \to 1$ as $\sigma^2 \to 0$.

Finally we have shown that $\Pr\{EEF(p) > EEF(j)\} \to 1$ for all $j \neq p$. Since $\Pr\{A_1 \cap A_2\} \to 1$ if $\Pr\{A_1\} \to 1$ and $\Pr\{A_2\} \to 1$ [12], as a result,

$$\Pr\left\{\arg\max_i EEF(i) = p\right\} \to 1$$

as $\sigma^2 \to 0$. This completes the proof that the EEF is consistent for the linear model.

B. Consistency of the EEF in General

In the general case, the signal $\mathbf{s}(\boldsymbol{\theta}_i)$ does not have to be a linear transformation of $\boldsymbol{\theta}_i$, and the noise $\mathbf{w}$ does not have to be Gaussian. To prove the consistency of the EEF in general, we first write the model in (1) as

$$\mathcal{H}_i : \mathbf{x} = \mathbf{s}_i(\boldsymbol{\theta}_i) + \sigma_n \mathbf{u} \tag{25}$$

where the $N \times 1$ signal $\mathbf{s}_i(\boldsymbol{\theta}_i)$ depends on the $i \times 1$ unknown parameters $\boldsymbol{\theta}_i$, and $\mathbf{u}$ has a well defined PDF and $\{\sigma_n\}$ is an arbitrary positive sequence that converges to 0. Because if we consider the probability of correct model order selection P_c as a function of σ^2, then the following conditions are equivalent [13]:

Condition 1)

$$\lim_{\sigma^2 \to 0} P_c(\sigma^2) = 1$$

Condition 2)

$$\lim_{n \to \infty} P_c(\sigma_n^2) = 1 \text{ for any arbitrary sequence } \{\sigma_n^2\} \text{ that converges to } 0$$

Hence we will prove Condition 2) to show the consistency of the EEF.

Let us assume the following.

Assumption 1): $\mathbf{s}(\boldsymbol{\theta}_i)$ is Lipschitz continuous, i.e., there exists $K > 0$ such that $\left\|\mathbf{s}_i(\boldsymbol{\theta}_i^1) - \mathbf{s}_i(\boldsymbol{\theta}_i^2)\right\| \leq K \left\|\boldsymbol{\theta}_i^1 - \boldsymbol{\theta}_i^2\right\|$ for all $\boldsymbol{\theta}_i^1$, $\boldsymbol{\theta}_i^2$.

Note that the linear signal $\mathbf{s}_i(\boldsymbol{\theta}_i) = \mathbf{H}_i\boldsymbol{\theta}_i$ is Lipschitz continuous since $\mathbf{s}_i(\boldsymbol{\theta}_i)$ is a linear transformation of $\boldsymbol{\theta}_i$ [14].

Assumption 2): The PDF $p_{\mathbf{U}}(\mathbf{u})$ of $\mathbf{u}$ satisfies

$$p_{\mathbf{U}}(\mathbf{u}_n)/p_{\mathbf{U}}(\mathbf{v}_n) \to \infty \text{ if } \|\mathbf{v}_n\| - \|\mathbf{u}_n\| \to \infty$$

and

$$\ln p_{\mathbf{U}}(\mathbf{u}) \text{ is Lipschitz continuous on set } \{\mathbf{u} : \|\mathbf{u}\| \leq l\} \text{ for any } l > 0$$

i.e., for any $l > 0$, there exists $L > 0$ such that $|\ln p_{\mathbf{U}}(\mathbf{u}_1) - \ln p_{\mathbf{U}}(\mathbf{u}_2)| \leq L\|\mathbf{u}_1 - \mathbf{u}_2\|$ for all $\mathbf{u}_1$, $\mathbf{u}_2$ with $\|\mathbf{u}_1\| \leq l$, $\|\mathbf{u}_2\| \leq l$.

Note that the Gaussian and Gaussian mixture PDFs will satisfy Assumption 2). For example, let the Gaussian mixture PDF be

$$p_{\mathbf{U}}(\mathbf{u}) = \sum_{i=1}^{m} \frac{\alpha_i}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{\|\mathbf{u}\|^2}{2\sigma_i^2}}$$

where $\alpha_i > 0$ and $\sum_{i=1}^m \alpha_i = 1$. Let $\sigma^2_{\max} = \max\{\sigma_1^2, \ldots, \sigma_m^2\}$, $\sigma^2_{\min} = \min\{\sigma_1^2, \ldots, \sigma_m^2\}$, and α be the α_i that corresponds to $\sigma^2_{\max}$. Then we have

$$\frac{p_{\mathbf{U}}(\mathbf{u})}{p_{\mathbf{U}}(\mathbf{v})} = \frac{\sum_{i=1}^{m} \frac{\alpha_i}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{\|\mathbf{u}\|^2}{2\sigma_i^2}}}{\sum_{i=1}^{m} \frac{\alpha_i}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{\|\mathbf{v}\|^2}{2\sigma_i^2}}} > \frac{\frac{\alpha}{\sqrt{2\pi\sigma^2_{\max}}} e^{-\frac{\|\mathbf{u}\|^2}{2\pi\sigma^2_{\max}}}}{\frac{1}{\sqrt{2\pi\sigma^2_{\min}}} e^{-\frac{\|\mathbf{v}\|^2}{2\sigma^2_{\max}}}} = \alpha\sqrt{\frac{\sigma^2_{\min}}{\sigma^2_{\max}}} \exp\left(\frac{\|\mathbf{v}\|^2 - \|\mathbf{u}\|^2}{2\sigma^2_{\max}}\right)$$

So if $\|\mathbf{v}_n\| - \|\mathbf{u}_n\| \to \infty$, it follows that $\|\mathbf{v}_n\|^2 - \|\mathbf{u}_n\|^2 \to \infty$ and hence $p_{\mathbf{U}}(\mathbf{u}_n)/p_{\mathbf{U}}(\mathbf{v}_n) \to \infty$.

Let $\mathcal{H}_p$ be the true model. With the above assumptions, the following theorems are proved in Appendices C-E.

Theorem 3 ($l_{G_j}(\mathbf{x})$ **unbounded in probability for** $j \geq p$)**.** *There exists a sequence* $\{N_n\}$ *with* $N_n \to \infty$ *such that* $\Pr\{l_{G_j}(\mathbf{x}) > N_n\} \to 1$ *as* $\sigma_n \to 0$ *for* $j \geq p$.

Note that each $\{N_n\}$ implicitly depends on σ_n. For example, in the linear model for $j \geq p$,

$$l_{G_j}(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{P}_j\mathbf{x}}{\sigma^2} \sim \chi_j'^2(\lambda)$$

where $\lambda = \frac{\left\|\mathbf{H}_p\boldsymbol{\theta}_p\right\|^2}{\sigma_n^2}$. If we choose $N_n = \frac{\left\|\mathbf{H}_p\boldsymbol{\theta}_p\right\|^2}{2\sigma_n^2}$, it can be shown that $\Pr\{l_{G_j}(\mathbf{x}) > N_n\} \to 1$ as $\sigma_n \to 0$.

Theorem 4 ($l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x})$ **bounded in probability for** $j > p$)**.** *For any sequence* $\{m_n\}$, $\Pr\{l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x}) < m_n\} \to 1$ *as* $m_n \to \infty$ *for* $j > p$.

Here the sequence $\{m_n\}$ can be an arbitrary sequence with $m_n \to \infty$, so m_n does not depend on σ_n. For example, in the linear model for $j > p$,

$$l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x}) \sim \chi^2_{j-p}$$

So for any $\{m_n\}$, $\Pr\{l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x}) < m_n\} \to 1$ as $m_n \to \infty$ for $j > p$.

Theorem 5 ($l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x})$ **unbounded in probability for** $j < p$)**.** *There exists a sequence* $\{M_n\}$ *with* $M_n \to \infty$ *such that* $\Pr\{l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x}) > M_n\} \to 1$ *as* $\sigma_n \to 0$ *for* $j < p$.

Note that each M_n also implicitly depends on σ_n. For example, in the linear model for $j < p$, by (20),

$$l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x}) \sim \chi'^2_{p-j}\left(\sum_{i=j}^{p-1} \alpha_i/\sigma_n^2\right)$$

If we choose $M_n = \sum_{i=j}^{p-1} \alpha_i/2\sigma_n^2$, it can be shown that $\Pr\{l_{G_j}(\mathbf{x}) > M_n\} \to 1$ as $\sigma_n \to 0$.

First we consider when $j > p$. For each σ_n, let $D_n^j = \{\mathbf{u} : l_{G_j}(\mathbf{x}) > N_n\}$, $D_n^p = \{\mathbf{u} : l_{G_p}(\mathbf{x}) > N_n\}$, $E_n = \{\mathbf{u} : l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x}) < m_n\}$, and $F_n = \{\mathbf{u} : EEF(p) > EEF(j)\}$. Then for any $\mathbf{u} \in D_n^j \cap D_n^p \cap E_n$, since $N_n \to \infty$, we can omit the unit function in the EEF. So we have

$$\begin{aligned} EEF(p) - EEF(j) &= l_{G_p}(\mathbf{x}) - p\left(\ln\frac{l_{G_p}(\mathbf{x})}{p} + 1\right) - l_{G_j}(\mathbf{x}) + j\left(\ln\frac{l_{G_j}(\mathbf{x})}{j} + 1\right) \\ &= p\ln\frac{l_{G_j}(\mathbf{x})}{l_{G_p}(\mathbf{x})} + (j-p)\ln l_{G_j}(\mathbf{x}) - \big(l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x})\big) + p\ln p - j\ln j - p + j \end{aligned} \tag{26}$$

Note that $\frac{l_{G_j}(\mathbf{x})}{l_{G_p}(\mathbf{x})} \geq 1$, $\ln l_{G_j}(\mathbf{x}) > \ln N_n$, and $l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x}) < m_n$. Since m_n is arbitrary, we can choose $m_n < (j-p)\ln N_n + p\ln p - j\ln j - p + j$ but still with $m_n \to \infty$ so that $EEF(p) - EEF(j) > 0$. This shows that $D_n^j \cap D_n^p \cap E_n \subseteq F_n$. By Theorems 3 and 4, we have $\Pr\{D_n^j\} \to 1$, $\Pr\{D_n^p\} \to 1$ and $\Pr\{E_n\} \to 1$, and hence $\Pr\{D_n^j \cap D_n^p \cap E_n\} \to 1$. This shows that $\Pr\{F_n\} \to 1$ as $\sigma_n \to 0$, i.e., $\Pr\{EEF(p) > EEF(j)\} \to 1$ as $\sigma_n \to 0$ for $j > p$.

Next, when $j < p$, let $D_n^p = \{\mathbf{u} : l_{G_p}(\mathbf{x}) > N_n\}$, $G_n = \{\mathbf{u} : l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x}) > M_n\}$, and $H_n = \{\mathbf{u} : EEF(p) > EEF(j)\}$ for each σ_n. Note that H_n and F_n are different since the former is for $j < p$ and the latter is for $j > p$. For any $\mathbf{u} \in D_n^k \cap G_n$, we have

$$EEF(p) - EEF(j) \;\; = \big(l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x})\big) + j\ln l_{G_j}(\mathbf{x}) - p\ln l_{G_p}(\mathbf{x}) + p\ln p - j\ln j - p + j \tag{27}$$

Since $x - p\ln x$ increases as x increases for $x > p$, we can find N_n and M_n such that $EEF(p) - EEF(j) > 0$. This shows that $D_n^p \cap G_n \in H_n$. By Theorem 3 with $j = p$ and Theorem 5, the rest of the proof is the same as for $j > p$.

Since we have shown that $\Pr\{EEF(p) > EEF(j)\} \to 1$ for all $j \neq p$, we have $\Pr\{\arg\max_i EEF(i) = p\} \to 1$ as $\sigma^2 \to 0$ using the property that $\Pr\{A_1 \cap A_2\} \to 1$ if $\Pr\{A_1\} \to 1$ and $\Pr\{A_2\} \to 1$ [12].

V. Simulation Results

A. Linear Signal

For the linear model when $M = 2$:

$$\mathcal{H}_1 : \mathbf{x} = \mathbf{h}_1\theta_1 + \mathbf{w}$$
$$\mathcal{H}_2 : \mathbf{x} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \mathbf{w} = \mathbf{H}_2\boldsymbol{\theta}_2 + \mathbf{w}$$

If $\mathcal{H}_1$ is true, by (4) and (5), the probability that the MDL will choose $\mathcal{H}_2$ is

$$\Pr\{\mathcal{H}_2|\mathcal{H}_1\} = \Pr\{MDL(1) \geq MDL(2)|\mathcal{H}_1\} = \Pr\{y_1 \geq \ln N|\mathcal{H}_1\} = 2Q\left(\sqrt{\ln N}\right) \tag{28}$$

So in this case, the lower bound is exactly the probability of overestimation error for the MDL. For the AIC, the lower bound $2Q\left(\sqrt{2}\right)$ is also exactly the probability of overestimation error. Hence the probabilities of correct model order selection P_c (note here that there is no underestimation error since the correct order is $k = 1$) for the MDL and the AIC are

$$P_c(MDL) = 1 - 2Q\left(\sqrt{\ln N}\right)$$
$$P_c(AIC) = 1 - 2Q\left(\sqrt{2}\right)$$

For the simulation, we use $N = 20$, $\mathbf{h}_1 = [1, 1, \ldots, 1]^T$, $\mathbf{h}_2 = [1, -1, 1, -1, \ldots, 1, -1]^T$, $\theta_1 = 1$ and $\theta_2 = 0$. We plot P_c versus $1/\sigma^2$. It can be expected that $P_c(MDL) = 1 - 2Q\left(\sqrt{\ln 20}\right) = 0.917$ and $P_c(AIC) = 1 - 2Q\left(\sqrt{2}\right) = 0.843$, and Figure 1 verifies our result. We can see that the EEF appears to be consistent in accordance with theorem, and the MDL and the AIC are inconsistent. Also the performances of the MDL and the AIC do not depend on σ^2.

Next we consider polynomial order estimation, which is essentially a linear model. We assume that $M = 4$, $N = 20$ and the true model order is $\mathcal{H}_3$ with the nth element of $\mathbf{s}(\boldsymbol{\theta}_3)$ being $s[n] = 0.1 + 0.3n + 0.1n^2$ for $n = 0, 1, \ldots, N-1$. P_c is plotted versus $1/\sigma^2$. As shown in Figure 2, the EEF is consistent and the MDL and the AIC are inconsistent. In this case, we cannot find P_c explicitly for the MDL and the AIC, but we can see that the performances of the MDL and the AIC are bounded above by $1 - 2Q\left(\sqrt{\ln 20}\right) = 0.917$ and $1 - 2Q\left(\sqrt{2}\right) = 0.843$ respectively.

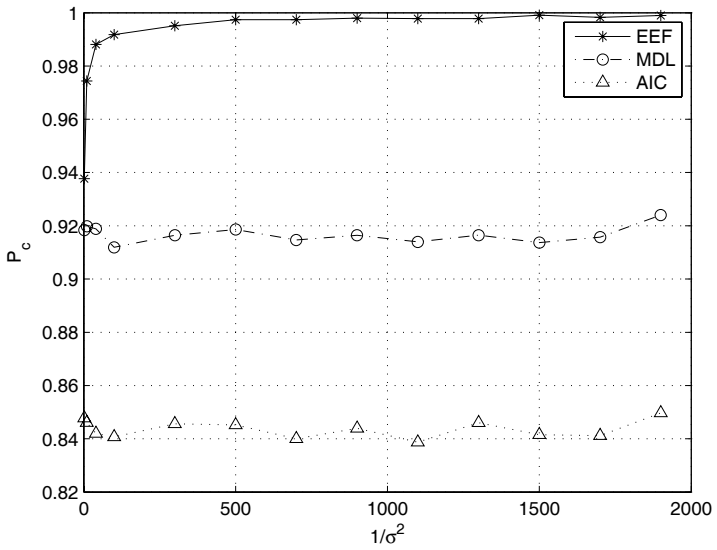

Fig. 1. Performance of MDL, AIC and EEF for the linear model when $\mathcal{H}_1$ is true (M=2, N=20).

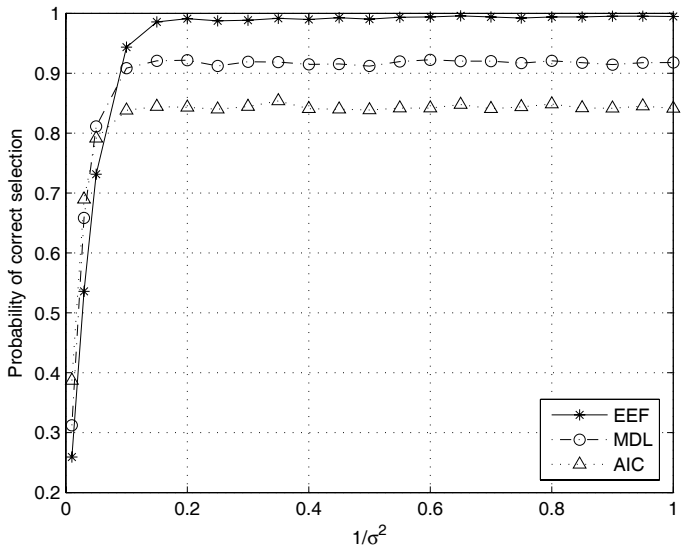

Fig. 2. Performance of MDL, AIC and EEF in estimating the polynomial model order when $\mathcal{H}_3$ is true (M=4, N=20).

B. Non-Linear Signal

We consider a problem of estimating of number of sinusoids. Suppose that under the ith model, the signal consists of i sinusoids embedded in white Gaussian noise. That is,

$$\mathcal{H}_i : x[n] = \sum_{j=1}^{i} A_j \cos\left(2\pi f_j n + \phi_j\right) + w[n]$$

for $n = 0, 1, \ldots, N-1$, $i = 1, 2, \ldots, M$, where the amplitudes A_j's, the frequencies f_j's and the phases ϕ_j's are unknown. To make the problem identifiable, we assume that $A_j > 0$, $0 < f_j < 1/2$, and $0 \leq \phi_j < 2\pi$. It can be easily checked that Assumptions 1) and 2) are satisfied for this example. Notice that if the frequencies f_j's are known, the model can be reduced to the linear model [11]

$$\mathcal{H}_i : \mathbf{x} = \mathbf{H}_i \boldsymbol{\alpha}_i + \mathbf{w} \tag{29}$$

where

$$\mathbf{H}_i = \begin{bmatrix} 1 & 0 & \cdots & 1 & 0 \\ \cos 2\pi f_1 & \sin 2\pi f_1 & \cdots & \cos 2\pi f_i & \sin 2\pi f_i \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \cos\left(2\pi f_1(N-1)\right) & \sin\left(2\pi f_1(N-1)\right) & \cdots & \cos\left(2\pi f_i(N-1)\right) & \sin\left(2\pi f_i(N-1)\right) \end{bmatrix}$$

is an $N \times 2i$ observation matrix for the ith model, and

$$\boldsymbol{\alpha}_i = \left[A_1 \cos\phi_1, -A_1 \sin\phi_1, \ldots, A_i \cos\phi_i, -A_i \sin\phi_i\right]^T$$

is a one-to-one transformation of the amplitudes A_j's and phases ϕ_j's. As a result, the MLEs of A_j's and ϕ_j's can be found from the MLE of $\boldsymbol{\alpha}_i$ according to the linear model in (29) whose observation matrix $\mathbf{H}_i$ depends on f_j's. So the MLE of $\boldsymbol{\alpha}_i$ is

$$\hat{\boldsymbol{\alpha}}_i = \left(\mathbf{H}_i^T \mathbf{H}_i\right)^{-1} \mathbf{H}_i^T \mathbf{x} \tag{30}$$

which is a function of f_j's for $j = 1, 2, \ldots, i$.

If the frequencies f_j's are unknown, as a result of (30), the MLEs of f_j's can be found by maximizing the following over the f_j's

$$g(f_1, f_2, \ldots, f_i) = \mathbf{x}^T \mathbf{H}_i \left(\mathbf{H}_i^T \mathbf{H}_i\right)^{-1} \mathbf{H}_i^T \mathbf{x} \tag{31}$$

Note that (31) is a function of f_j's because $\mathbf{H}_i$ depends on f_j's.

We denote the observation matrix $\mathbf{H}_i$ corresponding to the MLE of f_j's as $\hat{\mathbf{H}}_i$. Note that the number of unknown parameters is $3i$ under $\mathcal{H}_i$. Similar to the previous subsection, the MDL, the AIC and the

EEF choose the model order with the largest of the following respectively

$$
\begin{aligned}
-MDL(i) &= \frac{\mathbf{x}^T\hat{\mathbf{H}}_i\left(\hat{\mathbf{H}}_i^T\hat{\mathbf{H}}_i\right)^{-1}\hat{\mathbf{H}}_i^T\mathbf{x}}{\sigma^2} - 3i\ln N \\
-AIC(i) &= \frac{\mathbf{x}^T\hat{\mathbf{H}}_i\left(\hat{\mathbf{H}}_i^T\hat{\mathbf{H}}_i\right)^{-1}\hat{\mathbf{H}}_i^T\mathbf{x}}{\sigma^2} - 6i \\
&EEF(i) \\
&= \left(\frac{\mathbf{x}^T\hat{\mathbf{H}}_i\left(\hat{\mathbf{H}}_i^T\hat{\mathbf{H}}_i\right)^{-1}\hat{\mathbf{H}}_i^T\mathbf{x}}{\sigma^2} - 3i\left[\ln\left(\frac{\mathbf{x}^T\hat{\mathbf{H}}_i\left(\hat{\mathbf{H}}_i^T\hat{\mathbf{H}}_i\right)^{-1}\hat{\mathbf{H}}_i^T\mathbf{x}}{3i\sigma^2}\right)+1\right]\right) \\
&\cdot u\left(\frac{\mathbf{x}^T\hat{\mathbf{H}}_i\left(\hat{\mathbf{H}}_i^T\hat{\mathbf{H}}_i\right)^{-1}\hat{\mathbf{H}}_i^T\mathbf{x}}{3i\sigma^2} - 1\right)
\end{aligned}
\tag{32}
$$

In the simulation, we assume that $M = 3$, $N = 20$ and the true model order is $\mathcal{H}_2$ with $s[n] = cos(2\pi 0.1n) + 0.8cos(2\pi 0.3n + \pi/5)$ for $n = 0, 1, \ldots, N-1$. The MLEs of f_j's that maximizes (31) are found by grid search. In Figure 3, we also observe the consistency of the EEF and the inconsistency of the MDL and the AIC as $\sigma^2 \to 0$. The probabilities of correct selection appear to have upper bounds for the MDL and the AIC, although no explicit bounds are calculated in this non-linear signal case.

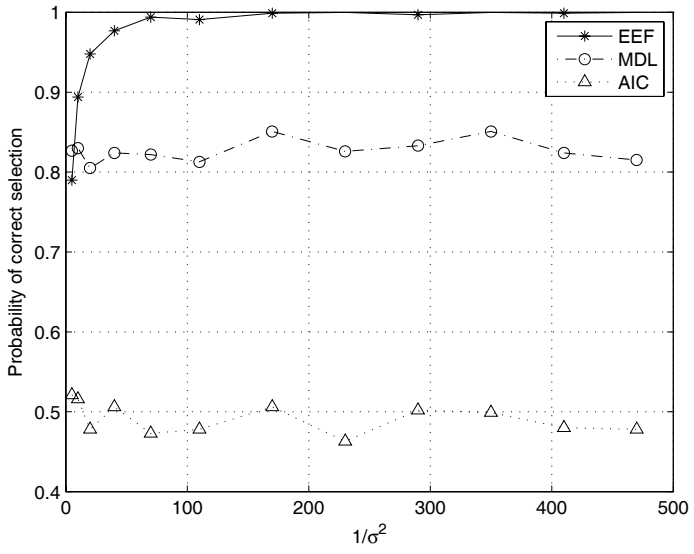

Fig. 3. Probability of correct selection for MDL, AIC and EEF in estimating the number of sinusoids when $\mathcal{H}_2$ is true (M=3, N=20).

VI. Conclusion

The inconsistency as $\sigma^2 \to 0$ of the MDL and the AIC has been shown. A simple lower bound is provided for their overestimating tendency. The consistency as $\sigma^2 \to 0$ of the EEF is also proved. Simulation results show that the EEF performs perfect under small noise while the MDL and the AIC do not.

Appendix A
Derivation of the Distribution of y_j's for $j \geq p$

We need the following lemma to derive the distribution of y_j's.

Lemma 1. $\mathbf{P}_{j+1} - \mathbf{P}_j$ *has rank 1.*

Proof: Suppose that for the subspace V_j generated by $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_j$, we have an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_j\}$. Then for the subspace V_{j+1} generated by $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{j+1}$, we can have an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_j, \mathbf{v}_{j+1}\}$. Since $\mathbf{P}_j$ is the projection matrix onto the subspace V_j, for any $N \times 1$ vector $\mathbf{x}$, we have

$$\mathbf{P}_j \mathbf{x} = \sum_{i=1}^{j} < \mathbf{x}, \mathbf{v}_i > \mathbf{v}_i \tag{33}$$

where $< \mathbf{x}, \mathbf{v_i} >$ is the inner product defined by

$$< \mathbf{x}, \mathbf{v}_i >= \mathbf{x}^T \mathbf{v}_i$$

Similarly, we also have

$$\mathbf{P}_{j+1} \mathbf{x} = \sum_{i=1}^{j+1} < \mathbf{x}, \mathbf{v}_i > \mathbf{v}_i \tag{34}$$

So (33) and (34) tell us that for any $\mathbf{x}$,

$$(\mathbf{P}_{j+1} - \mathbf{P}_j)\, \mathbf{x} =< \mathbf{x}, \mathbf{v}_{j+1} > \mathbf{v}_{j+1} = \alpha \mathbf{v}_{j+1} \tag{35}$$

for a scalar α. This shows that $\mathbf{P}_{j+1} - \mathbf{P}_j$ has rank 1 since it projects any $\mathbf{x}$ onto the 1-dimensional subspace generated by $\mathbf{v}_{j+1}$. ■

Since we assume under $\mathcal{H}_p$ that $\mathbf{x} = \mathbf{H}_p \boldsymbol{\theta}_p + \mathbf{w}$,

$$y_p = \frac{(\mathbf{H}_p \boldsymbol{\theta}_p + \mathbf{w})^T (\mathbf{P}_{p+1} - \mathbf{P}_p) (\mathbf{H}_p \boldsymbol{\theta}_p + \mathbf{w})}{\sigma^2}. \tag{36}$$

Since $\mathbf{H}_p \boldsymbol{\theta}_p = \sum_{i=1}^{p} \theta_i \mathbf{h}_i \in V_p$, the projection of $\mathbf{H}_p \boldsymbol{\theta}_p$ onto V_p remains the same. That is,

$$\mathbf{P}_p \mathbf{H}_p \boldsymbol{\theta}_p = \mathbf{H}_p \boldsymbol{\theta}_p.$$

Also $\mathbf{H}_p\boldsymbol{\theta}_p = \sum_{i=1}^{p} \theta_i \mathbf{h}_i + 0\mathbf{h}_{p+1} \in V_{p+1}$, thus $\mathbf{P}_{p+1}\mathbf{H}_p\boldsymbol{\theta}_p = \mathbf{H}_p\boldsymbol{\theta}_p$. So we have

$$(\mathbf{P}_{p+1} - \mathbf{P}_p)\,\mathbf{H}_p\boldsymbol{\theta}_p = \mathbf{0}$$

and hence

$$y_p = \frac{\mathbf{w}^T(\mathbf{P}_{p+1} - \mathbf{P}_p)\,\mathbf{w}}{\sigma^2} = \mathbf{u}^T(\mathbf{P}_{p+1} - \mathbf{P}_p)\,\mathbf{u} \tag{37}$$

where $\mathbf{u} = \mathbf{w}/\sigma$ is an $N \times 1$ white Gaussian noise vector with unit variance.

For $j > p$, we can think of $\mathbf{H}_p\boldsymbol{\theta}_p$ as $\mathbf{H}_j\boldsymbol{\theta}_j$ where $\boldsymbol{\theta}_j = [\theta_1, \theta_2, \ldots, \theta_p, 0, \ldots, 0]^T$. By the same derivation as above, we can also show that

$$y_j = \mathbf{u}^T(\mathbf{P}_{j+1} - \mathbf{P}_j)\,\mathbf{u}. \tag{38}$$

It is well known that $\mathbf{P}_j$ is a symmetric idempotent matrix and $\mathbf{P}_{j+1}\mathbf{P}_j = \mathbf{P}_j$. So

$$(\mathbf{P}_{j+1} - \mathbf{P}_j)(\mathbf{P}_{j+1} - \mathbf{P}_j) = \mathbf{P}_{j+1} - \mathbf{P}_j.$$

This says that $\mathbf{P}_{j+1} - \mathbf{P}_j$ is also idempotent. By Lemma 1 $\mathbf{P}_{j+1} - \mathbf{P}_j$ has rank 1, so by [1]

$$y_j = \mathbf{u}^T(\mathbf{P}_{j+1} - \mathbf{P}_j)\,\mathbf{u} \sim \chi_1^2 \quad \text{for all } j \geq p. \tag{39}$$

where χ_1^2 is the chi-square distribution with 1 degree of freedom.

We still need to show the independence of y_j's for all $j \geq p$. Let $\mathbf{z}_j = (\mathbf{P}_{j+1} - \mathbf{P}_j)\,\mathbf{u}$. Since $\mathbf{z}_j$ is a linear transform of $\mathbf{u}$, $\mathbf{z}_j$ is also Gaussian with zero mean. For any $l > 0$, we will show next that $\mathbf{z}_j$ and $\mathbf{z}_{j+l}$ are independent for any $j \geq p$.

Let $\begin{bmatrix} \mathbf{z}_j \\ \mathbf{z}_{j+l} \end{bmatrix} = \begin{bmatrix} \mathbf{P}_{j+1} - \mathbf{P}_j \\ \mathbf{P}_{j+l+1} - \mathbf{P}_{j+l} \end{bmatrix} \mathbf{u}$, whose covariance matrix is

$$\begin{aligned}
\mathbf{C}_{\mathbf{z}_j,\mathbf{z}_{+l}} &= \begin{bmatrix} \mathbf{P}_{j+1} - \mathbf{P}_j \\ \mathbf{P}_{j+l+1} - \mathbf{P}_{j+l} \end{bmatrix} \begin{bmatrix} \mathbf{P}_{j+1} - \mathbf{P}_j & \mathbf{P}_{j+l+1} - \mathbf{P}_{j+l} \end{bmatrix} \\
&= \begin{bmatrix} (\mathbf{P}_{j+1} - \mathbf{P}_j)(\mathbf{P}_{j+1} - \mathbf{P}_j) & (\mathbf{P}_{j+1} - \mathbf{P}_j)(\mathbf{P}_{j+l+1} - \mathbf{P}_{j+l}) \\ (\mathbf{P}_{j+l+1} - \mathbf{P}_{j+l})(\mathbf{P}_{j+1} - \mathbf{P}_j) & (\mathbf{P}_{j+l+1} - \mathbf{P}_{j+l})(\mathbf{P}_{j+l+1} - \mathbf{P}_{j+l}) \end{bmatrix}.
\end{aligned}$$

By the property of $\mathbf{P}_j$ that $\mathbf{P}_m\mathbf{P}_{m+n} = \mathbf{P}_m$ for $n > 0$, we have

$$\begin{aligned}
(\mathbf{P}_{j+1} - \mathbf{P}_j)(\mathbf{P}_{j+l+1} - \mathbf{P}_{j+l}) &= \mathbf{P}_{j+1}\mathbf{P}_{j+l+1} - \mathbf{P}_j\mathbf{P}_{j+l+1} - \mathbf{P}_{j+1}\mathbf{P}_{j+l} + \mathbf{P}_j\mathbf{P}_{j+l} \\
&= \mathbf{P}_{j+1} - \mathbf{P}_j - \mathbf{P}_{j+1} + \mathbf{P}_j \\
&= \mathbf{0}_{N\times N}.
\end{aligned}$$

This shows that $\mathbf{z}_j$ and $\mathbf{z}_{j+l}$ are uncorrelated and hence independent by Gaussianity. Also by Gaussianity, pairwise independence will lead to the independence of all z_j's. Since $y_j = \mathbf{z}_j^T \mathbf{z}_j$, we can say y_j's are independent $j \geq p$.

APPENDIX B
DERIVATION OF THE DISTRIBUTION OF y_j'S FOR $j < p$

If $\mathcal{H}_p$ is true, for $j < p$ we still have

$$y_j = \frac{(\mathbf{H}_p \boldsymbol{\theta}_p + \mathbf{w})^T (\mathbf{P}_{j+1} - \mathbf{P}_j)(\mathbf{H}_p \boldsymbol{\theta}_p + \mathbf{w})}{\sigma^2}. \tag{40}$$

But when $j < p$,

$$(\mathbf{P}_{j+1} - \mathbf{P}_j)\mathbf{H}_p \boldsymbol{\theta}_p \neq \mathbf{0}$$

so we cannot reduce (40) as in (38). However, we can write y_j as

$$\begin{aligned} y_j &= \left(\frac{\mathbf{H}_p \boldsymbol{\theta}_p}{\sigma} + \mathbf{u}\right)^T (\mathbf{P}_{j+1} - \mathbf{P}_j) \left(\frac{\mathbf{H}_p \boldsymbol{\theta}_p}{\sigma} + \mathbf{u}\right) \\ &= \left(\frac{(\mathbf{P}_{j+1} - \mathbf{P}_j)\mathbf{H}_p \boldsymbol{\theta}_p}{\sigma} + \mathbf{z}_j\right)^T \left(\frac{(\mathbf{P}_{j+1} - \mathbf{P}_j)\mathbf{H}_p \boldsymbol{\theta}_p}{\sigma} + \mathbf{z}_j\right) \end{aligned} \tag{41}$$

where $\mathbf{u} = \mathbf{w}/\sigma$ and $\mathbf{z}_j = (\mathbf{P}_{j+1} - \mathbf{P}_j)\mathbf{u}$ as in Appendix A. Since we have shown that $\mathbf{z}_j^T \mathbf{z}_j \sim \chi_1^2$, we have

$$y_j \sim \chi_1^{'2}(\lambda_j) \tag{42}$$

where $\chi_1^{'2}(\lambda_j)$ is the noncentral chi-square distribution with 1 degree of freedom and noncentrality parameter $\lambda_j = \|(\mathbf{P}_{j+1} - \mathbf{P}_j)\mathbf{H}_p \boldsymbol{\theta}_p\|^2 / \sigma^2 = (\mathbf{H}_p \boldsymbol{\theta}_p)^T (\mathbf{P}_{j+1} - \mathbf{P}_j)\mathbf{H}_p \boldsymbol{\theta}_p / \sigma^2 > 0$. If we let $\mathbf{H}_{j+1,p} = [\mathbf{h}_{j+1}, \mathbf{h}_{j+2}, \ldots, \mathbf{h}_p]$ and $\boldsymbol{\theta}_{j+1,p} = [\theta_{j+1}, \theta_{j+2}, \ldots, \theta_p]^T$, since $(\mathbf{P}_{j+1} - \mathbf{P}_j)\mathbf{H}_j \boldsymbol{\theta}_j = \mathbf{0}$, we have

$$\begin{aligned} \lambda_j &= (\mathbf{H}_j \boldsymbol{\theta}_j + \mathbf{H}_{j+1,p} \boldsymbol{\theta}_{j+1,p})^T (\mathbf{P}_{j+1} - \mathbf{P}_j)(\mathbf{H}_j \boldsymbol{\theta}_j + \mathbf{H}_{j+1,p} \boldsymbol{\theta}_{j+1,p}) / \sigma^2 \\ &= (\mathbf{H}_{j+1,p} \boldsymbol{\theta}_{j+1,p})^T (\mathbf{P}_{j+1} - \mathbf{P}_j)\mathbf{H}_{j+1,p} \boldsymbol{\theta}_{j+1,p} / \sigma^2 \end{aligned}$$

So λ_j does not depend on the first j θ_i's in $\boldsymbol{\theta}_p$.

Since the proof of the independence of $\mathbf{z}_j$'s in Appendix A does not depend on whether $j \geq p$ or $j < p$, $\mathbf{z}_j$'s are independent for all j. Hence so are y_j's.

APPENDIX C
PROOF OF THEOREM 3

Theorem 3 ($l_{G_j}(\mathbf{x})$ **unbounded in probability for** $j \geq p$). There exists a sequence $\{N_n\}$ with $N_n \to \infty$ such that $\Pr\{l_{G_j}(\mathbf{x}) > N_n\} \to 1$ as $\sigma_n \to 0$ for $j \geq p$.

First we will prove the next lemma.

Lemma 2. *Under the true model,* $\mathbf{s}_p(\hat{\boldsymbol{\theta}}_p) \xrightarrow{P} \mathbf{s}_p(\boldsymbol{\theta}_p)$ *as* $\sigma_n \to 0$. *That is, for any* $\epsilon > 0$, $\Pr\left\{\left\|\mathbf{s}_p\left(\hat{\boldsymbol{\theta}}_p\right) - \mathbf{s}_p(\boldsymbol{\theta}_p)\right\| < \epsilon\right\} \to 1$ *as* $\sigma_n \to 0$.

Proof: First we will introduce the work in [15], which considers the characteristics of the MLE under high SNR. Let

$$\mathbf{f}(\boldsymbol{\theta}_p, \mathbf{u}) = [f_1(\boldsymbol{\theta}_p, \mathbf{u}), \ldots, f_p(\boldsymbol{\theta}_p, \mathbf{u})]^T = \frac{\partial p_{\mathbf{U}}\left(\frac{\mathbf{x}(\mathbf{u}) - \mathbf{s_p}(\boldsymbol{\theta}_\mathbf{p})}{\sigma_n}\right)}{\partial \boldsymbol{\theta}_p}$$

where we consider $\mathbf{x}$ is a function of $\mathbf{u}$, then the MLE of $\boldsymbol{\theta}_p$ is found by solving

$$\mathbf{f}(\boldsymbol{\theta}_p, \mathbf{u}) = \mathbf{0}$$

If $f_i(\boldsymbol{\theta}_p, \mathbf{u})$ for $i = 1, \ldots, p$ are differentiable functions on a neighborhood of a point $(\boldsymbol{\theta}_p^0, \mathbf{u}_0)$ with $\mathbf{f}(\boldsymbol{\theta}_p^0, \mathbf{u}_0) = \mathbf{0}$, and the Jacobian matrix $\boldsymbol{\Phi}$ with respect to $\mathbf{u}$ is nonsingular at $(\boldsymbol{\theta}_p^0, \mathbf{u}_0)$, then by the implicit function theorem, we have

$$\frac{\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_p}{\sigma_n} \xrightarrow{P} -\boldsymbol{\Phi}^{-1}\boldsymbol{\Psi}\mathbf{u} \tag{43}$$

where $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are determined matrices with

$$\boldsymbol{\Phi} = \left[\frac{\partial \mathbf{f}}{\partial u_1}\Big|_{(\boldsymbol{\theta}_p^0, \mathbf{u}_0)}, \ldots, \frac{\partial \mathbf{f}}{\partial u_N}\Big|_{(\boldsymbol{\theta}_p^0, \mathbf{u}_0)}\right]$$
$$\boldsymbol{\Psi} = \left[\frac{\partial \mathbf{f}}{\partial \theta_1}\Big|_{(\boldsymbol{\theta}_p^0, \mathbf{u}_0)}, \ldots, \frac{\partial \mathbf{f}}{\partial \theta_p}\Big|_{(\boldsymbol{\theta}_p^0, \mathbf{u}_0)}\right]$$

Although only Gaussian noise is considered in [15], (43) still holds for non-Gaussian noise by the implicit function theorem.

It has been shown in [12] that if $\{\mathbf{X}_n\}$ is a sequence of random variables that converges to $\mathbf{X}$ in probability and $\{c_n\}$ is a determined sequence that converges to c, then $c_n\mathbf{X}_n \xrightarrow{P} c\mathbf{X}$. As a result of (43), since $\sigma_n \to 0$, we have

$$\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_p = \sigma_n \frac{\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_p}{\sigma_n} \xrightarrow{P} 0 \tag{44}$$

Then by Assumption 1), $\left\|\mathbf{s}_p\left(\hat{\boldsymbol{\theta}}_p\right) - \mathbf{s}_p(\boldsymbol{\theta}_p)\right\| \xrightarrow{P} 0$. This completes the proof of Lemma 2. ■

When the true model is $\mathcal{H}_p$, for $j > p$, the MLE for $\boldsymbol{\theta}_j$ is still under the true model if we write $\boldsymbol{\theta}_j$ as $\boldsymbol{\theta}_j = [\boldsymbol{\theta}_p^T, 0, \ldots, 0]^T$. So from (44), we have $\hat{\boldsymbol{\theta}}_j \xrightarrow{P} \boldsymbol{\theta}_j$, i.e.,

$$\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \vdots \\ \vdots \\ \hat{\theta}_j \end{bmatrix} \xrightarrow{P} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Hence Lemma 2 still holds for $j > p$, and it extends to

$$\left\| \mathbf{s}_j\left(\hat{\boldsymbol{\theta}}_j\right) - \mathbf{s}_j\left(\boldsymbol{\theta}_j\right) \right\| \xrightarrow{P} 0 \text{ for all } j \geq p \tag{45}$$

So we have

$$\begin{aligned} l_{G_j}(\mathbf{x}) &= 2\ln \frac{p_{\mathbf{U}}\left(\frac{\mathbf{x}-\mathbf{s}_j(\hat{\boldsymbol{\theta}}_j)}{\sigma_n}\right)\frac{1}{\sigma_n}}{p_{\mathbf{U}}\left(\frac{\mathbf{x}}{\sigma_n}\right)\frac{1}{\sigma_n}} \\ &= 2\ln \frac{p_{\mathbf{U}}\left(\frac{\mathbf{s}_j(\boldsymbol{\theta}_j)+\sigma_n\mathbf{u}-\mathbf{s}_j(\hat{\boldsymbol{\theta}}_j)}{\sigma_n}\right)}{p_{\mathbf{U}}\left(\frac{\mathbf{s}_j(\boldsymbol{\theta}_j)+\sigma_n\mathbf{u}}{\sigma_n}\right)} \end{aligned} \tag{46}$$

Since $p_{\mathbf{U}}(\mathbf{u})$ is a well defined PDF, we have

$$\Pr\{\|\mathbf{u}\| < l_n\} \to 1 \tag{47}$$

for any sequence $\{l_n\}$ with $l_n \to \infty$.

Let $A_n = \{\mathbf{u} : \left\| \mathbf{s}_j\left(\hat{\boldsymbol{\theta}}_j\right) - \mathbf{s}_j\left(\boldsymbol{\theta}_j\right) \right\| < \epsilon\}$ and $B_n = \{\mathbf{u} : \|\mathbf{u}\| < l_n\}$ for each σ_n. Since l_n and ϵ are arbitrary, we let $l_n = \|\mathbf{s}_j(\boldsymbol{\theta}_j)\|/(3\sigma_n)$ and $\epsilon = \|\mathbf{s}_j(\boldsymbol{\theta}_j)\|/6$. Then for each $\mathbf{u} \in A_n \cap B_n$, we have

$$\frac{\left\|\mathbf{s}_j(\boldsymbol{\theta}_j) + \sigma_n\mathbf{u} - \mathbf{s}_j(\hat{\boldsymbol{\theta}}_j)\right\|}{\sigma_n} \leq \frac{\left\|\mathbf{s}_j(\boldsymbol{\theta}_j) - \mathbf{s}_j(\hat{\boldsymbol{\theta}}_j)\right\|}{\sigma_n} + \|\mathbf{u}\| < \frac{\epsilon}{\sigma_n} + l_n \tag{48}$$

Hence

$$\begin{aligned} &\frac{\|\mathbf{s}_j(\boldsymbol{\theta}_j) + \sigma_n\mathbf{u}\|}{\sigma_n} - \frac{\left\|\mathbf{s}_j(\boldsymbol{\theta}_j) + \sigma_n\mathbf{u} - \mathbf{s}_j(\hat{\boldsymbol{\theta}}_j)\right\|}{\sigma_n} \\ &> \left(\frac{\|\mathbf{s}_j(\boldsymbol{\theta}_j)\|}{\sigma_n} - \|\mathbf{u}\|\right) - \left(\frac{\epsilon}{\sigma_n} + l_n\right) \\ &> \frac{\|\mathbf{s}_j(\boldsymbol{\theta}_j)\|}{\sigma_n} - 2l_n - \frac{\epsilon}{\sigma_n} \\ &= \frac{\|\mathbf{s}_j(\boldsymbol{\theta}_j)\|}{6\sigma_n} \to \infty \end{aligned} \tag{49}$$

as $\sigma_n \to 0$. By Assumption 2), this shows that $l_{G_j}(\mathbf{x}) \to \infty$ as $\sigma_n \to 0$ for each $\mathbf{u} \in A_n \cap B_n$. Let $C = \{\mathbf{u} : l_{G_j}(\mathbf{x}) \to \infty \text{ as } \sigma_n \to 0\}$. The previous analysis shows that $A_n \cap B_n \subseteq C$. By (45) and (47), $\Pr\{A_n\} \to 1$ and $\Pr\{B_n\} \to 1$ as $\sigma_n \to 0$. Hence $\Pr\{A_n \cap B_n\} \to 1$. Note that $A_n \cap B_n \subseteq C$, and thus $\Pr\{C\} = 1$. From this *"almost sure"* event, it follows the *"in probability"* event, i.e., for any $\epsilon > 0$ and any M, there exists an integer K such that $\Pr\{l_{G_j}(\mathbf{x}) \leq M\} < \epsilon$ for all $n \geq K$. Next, the existence of a sequence $\{N_n\}$ with $N_n \to \infty$ such that $\Pr\{l_{G_j}(\mathbf{x}) > N_n\} \to 1$ as $\sigma_n \to 0$ for $j \geq p$ will be shown by constructing such a sequence $\{N_n\}$.

Let $\{M_m\}$ be any sequence that goes to ∞. For each M_m, there exists K_m such that $\Pr\{l_{G_j}(\mathbf{x}) \leq M_m\} < \epsilon$ for all $n \geq K_m$. We construct $\{N_n\}$ as

$$\begin{array}{lccccccc} \{N_n\} = & 0 & , \ldots, 0, & M_1 & , \ldots, M_1, & M_2 & , \ldots \\ & \uparrow & & \uparrow & & \uparrow & \\ & \text{1st term} & & K_1\text{th term} & & K_2\text{th term} & \end{array}$$

So $N_n \to \infty$ since $M_m \to \infty$. For any n, we can find a m such that $K_m \leq n < K_{m+1}$, and $N_n = M_m$ by the above construction of $\{N_n\}$. Hence $\Pr\{l_{G_j}(\mathbf{x}) \leq N_n\} = \Pr\{l_{G_j}(\mathbf{x}) \leq M_m\} < \epsilon$ for all n. This proves the existence of a sequence $\{N_n\}$ with $N_n \to \infty$ such that $\Pr\{l_{G_j}(\mathbf{x}) > N_n\} \to 1$ as $\sigma_n \to 0$ for $j \geq p$.

APPENDIX D
PROOF OF THEOREM 4

Theorem 4 ($l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x})$ **bounded in probability for** $j > p$). For any sequence $\{m_n\}$, $\Pr\{l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x}) < m_n\} \to 1$ as $m_n \to \infty$ for $j > p$.

For $j > p$,

$$l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x}) = 2 \ln p_{\mathbf{U}} \left(\frac{\mathbf{s}_j(\boldsymbol{\theta}_j) + \sigma_n \mathbf{u} - \mathbf{s}_j(\hat{\boldsymbol{\theta}}_j)}{\sigma_n} \right) - 2 \ln p_{\mathbf{U}} \left(\frac{\mathbf{s}_p(\boldsymbol{\theta}_p) + \sigma_n \mathbf{u} - \mathbf{s}_p(\hat{\boldsymbol{\theta}}_p)}{\sigma_n} \right) \tag{50}$$

Note that we can consider $\boldsymbol{\theta}_j$ as $\boldsymbol{\theta}_j = [\boldsymbol{\theta}_p^T, 0, \ldots, 0]^T$, and so we have $\mathbf{s}_j(\boldsymbol{\theta}_j) = \mathbf{s}_p(\boldsymbol{\theta}_p)$.

By (43) and Assumption 1),

$$\frac{\left\| \mathbf{s}_j(\hat{\boldsymbol{\theta}}_j) - \mathbf{s}_p(\hat{\boldsymbol{\theta}}_p) \right\|}{\sigma_n} \leq \frac{\left\| \mathbf{s}_j(\boldsymbol{\theta}_j) - \mathbf{s}_j(\hat{\boldsymbol{\theta}}_j) \right\|}{\sigma_n} + \frac{\left\| \mathbf{s}_p(\boldsymbol{\theta}_p) - \mathbf{s}_p(\hat{\boldsymbol{\theta}}_p) \right\|}{\sigma_n} \leq K \frac{\left\| \boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j \right\|}{\sigma_n} + K \frac{\left\| \boldsymbol{\theta}_p - \hat{\boldsymbol{\theta}}_p \right\|}{\sigma_n} \xrightarrow{P} 2K \left\| \boldsymbol{\Phi}^{-1} \boldsymbol{\Psi} \mathbf{u} \right\| \tag{51}$$

By the Lipschitz continuity of $\ln p_{\mathbf{U}}(\mathbf{u})$, there exists L such that (50) can be written as

$$l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x}) = \left| 2\ln p_{\mathbf{U}}\left(\frac{\mathbf{s}_j(\boldsymbol{\theta}_j) + \sigma_n\mathbf{u} - \mathbf{s}_j(\hat{\boldsymbol{\theta}}_j)}{\sigma_n}\right) - 2\ln p_{\mathbf{U}}\left(\frac{\mathbf{s}_p(\boldsymbol{\theta}_p) + \sigma_n\mathbf{u} - \mathbf{s}_p(\hat{\boldsymbol{\theta}}_p)}{\sigma_n}\right)\right|$$
$$\leq 2L\frac{\left\|\mathbf{s}_j(\hat{\boldsymbol{\theta}}_j) - \mathbf{s}_p(\hat{\boldsymbol{\theta}}_p)\right\|}{\sigma_n} \leq K\frac{\left\|\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j\right\|}{\sigma_n} + K\frac{\left\|\boldsymbol{\theta}_p - \hat{\boldsymbol{\theta}}_p\right\|}{\sigma_n} \xrightarrow{P} 4LK\left\|\boldsymbol{\Phi}^{-1}\boldsymbol{\Psi}\mathbf{u}\right\| \quad (52)$$

where the second inequality is by (51). Similar to (47), we have

$$\Pr\{\left\|\boldsymbol{\Phi}^{-1}\boldsymbol{\Psi}\mathbf{u}\right\| < l_n\} \to 1 \quad (53)$$

and hence

$$\Pr\left\{l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x}) < 4LKl_n\right\} \to 1 \quad (54)$$

as $l_n \to \infty$. Since $\{l_n\}$ is an arbitrary sequence with $l_n \to \infty$, we have $\Pr\{l_{G_j}(\mathbf{x}) - l_{G_p}(\mathbf{x}) < m_n\} \to 1$ as $m_n \to \infty$ for any sequence $\{m_n\}$.

APPENDIX E
PROOF OF THEOREM 5

Theorem 5 ($l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x})$ **unbounded in probability for** $j < p$). There exists a sequence $\{M_n\}$ with $M_n \to \infty$ such that $\Pr\{l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x}) > M_n\} \to 1$ as $\sigma_n \to 0$ for $j < p$.

For $j < p$,

$$l_{G_p}(\mathbf{x}) - l_{G_j}(\mathbf{x}) = 2\ln p_{\mathbf{U}}\left(\frac{\mathbf{s}_p(\boldsymbol{\theta}_p) + \sigma_n\mathbf{u} - \mathbf{s}_p(\hat{\boldsymbol{\theta}}_p)}{\sigma_n}\right) - 2\ln p_{\mathbf{U}}\left(\frac{\mathbf{s}_p(\boldsymbol{\theta}_p) + \sigma_n\mathbf{u} - \mathbf{s}_j(\hat{\boldsymbol{\theta}}_j)}{\sigma_n}\right) \quad (55)$$

Note that we do not have $\mathbf{s}_j(\boldsymbol{\theta}_j) = \mathbf{s}_p(\boldsymbol{\theta}_p)$ as for the $j > p$ case, because it is under the misspecified model when $j < p$. This means that we cannot find $\boldsymbol{\theta}_j$ such that $\mathbf{s}_j(\boldsymbol{\theta}_j) = \mathbf{s}_p(\boldsymbol{\theta}_p)$ or $\mathbf{s}_j(\boldsymbol{\theta}_j)$ is arbitrarily close to $\mathbf{s}_p(\boldsymbol{\theta}_p)$. So we assume that there exists $\delta > 0$ such that $\|\mathbf{s}_j(\boldsymbol{\theta}_j) - \mathbf{s}_p(\boldsymbol{\theta}_p)\| > \delta$ for all $\boldsymbol{\theta}_j$. Then the rest of the proof follows similarly to the proof of Theorem 3 in Appendix C using Lemma 2 and Assumption 2).

REFERENCES

[1] S. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice-Hall, 1998.

[2] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, pp. 716–723, Dec. 1974.

[3] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[4] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[5] S. Kay, "Exponentially embedded families - new approaches to model order estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 41, pp. 333–345, Jan. 2005.

[6] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, pp. 36–47, 2004.

[7] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 387–392, Apr. 1985.

[8] C. Xu and S. Kay, "Source enumeration via the eef criterion," *IEEE Signal Process. Lett.*, vol. 15, pp. 569–572, 2008.

[9] R.A. Fisher, "On the mathematical foundations of theoretical statistics," *Philos. Trans. Royal Soc. London*, vol. 222, no. 594-604, pp. 309–368, 1922.

[10] R.E. Kass and P.W. Vos, *Geometrical Foundations of Asymptotic Inference*, Wiley, 1997.

[11] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, 1993.

[12] E.L. Lehmann, *Elements of Large-Sample Theory*, Springer, 1998.

[13] W. Rudin, *Functional Analysis*, McGraw-Hill, 1991.

[14] K. Eriksson, D.J. Estep, and C. Johnson, *Applied Mathematics, Body and Soul: Calculus in Several Dimensions*, Springer, 2004.

[15] A. Renaux, P. Forster, E. Chaumette, and P. Larzabal, "On the high-snr conditional maximum-likelihood estimator full statistical characterization," *IEEE Trans. Signal Process.*, vol. 54, pp. 4840–4843, 2006.