The Penalty Term of Exponentially Embedded Family is Estimated Mutual Information

Zhenghan Zhu, *Student member, IEEE*, and Steven Kay, *Life Fellow, IEEE*
Department of Electrical, Computer and Biomedical Engineering
University of Rhode Island, Kingston, RI, 02881
Email: zzhu@ele.uri.edu; kay@ele.uri.edu

***Abstract*—The penalty term plays an important role in model order selection rules. The Exponentially Embedded Families (EEF) is consistent and effective in model order selection. In this paper we show that the EEF penalty term can be viewed as estimated mutual information (MI) between unknown parameters and received data from Bayesian viewpoints. The finding is a result of an important relationship between Kullback-Leibler Divergence (KLD), signal-to-noise ratio (SNR) and MI in estimation/detection of random signals, which is also introduced.**

I. Introduction

Model order selection is a fundamental problem in signal processing because observed data in practice usually is composed of an unknown number of signal components. For example, one may need to determine the number of sources in array signal processing [1]. Overestimating the order actually fits the noise in the data; underestimating the order on the other hand fails to describe the data precisely [1].

Model order selection problem, as a multiple hypotheses testing problem, lacks an optimal solution [11]. The traditional generalized likelihood ratio test (GLRT) tends to overestimate the order [7]. As a result, a typical model order selection algorithm introduces a penalty term to form a decision rule. Several popular algorithms are Akaike's information criterion (AIC) [2], the minimum description length (MDL) [3], Bayesian information criterion (BIC) [4] and maximum a posteriori (MAP) [11]. The reference [5] provides a review in this regard.

In addition to the aforementioned rules, EEF has been introduced in [8] as an alternative. It embeds two PDFs into a family of PDFs that are indexed by one or more parameters, and the new embedded family inherits many mathematical and optimality properties of the exponential family. It proves effective in model order selection and even superior under certain conditions. It has been shown to be consistent, i.e., as the data length $N \to \infty$, the probability of selecting the correct model goes to one [1]. The penalty term plays a central role in the EEF model order selection algorithm. In this paper we show that the EEF penalty term is actually the estimated mutual information between the unknown parameters and the received data. This hopefully can shed further light to understanding in choosing optimal penalty term for model order selection. We limit the discussion in the context of linear normal model. A more general discussion will be our future work.

The paper is organized as follows. In Section II we introduce an useful relationship between KLD, SNR and MI, which holds in general in estimation/detection of random signals. In Section III a brief introduction is given to EEF. In Section IV we discuss the EEF penalty term with an illustrative example. We then extend the discussion to the linear model in Section V. Finally, some conclusions are drawn in Section VI.

II. An important relationship among KLD, SNR and MI

In signal processing, we often encounter problems of estimation/detection of random signals. Suppose we want to decide between the following hypotheses

$$\begin{aligned}\mathcal{H}_0 : \mathbf{x} &= \mathbf{w} \\ \mathcal{H}_1 : \mathbf{x} &= \mathbf{t} + \mathbf{w}\end{aligned}$$

where $\mathbf{w}$ is noise and $\mathbf{t}$ is a random signal. Denote $p_1(\mathbf{x})$ and $p_0(\mathbf{x})$ as the probability density function (PDF) of the received data $\mathbf{x}$ under $\mathcal{H}_1$ and $\mathcal{H}_0$ respectively, and $\pi(\mathbf{t})$ as the prior PDF of $\mathbf{t}$. An interesting and useful relationship is [6]

$$D(p_1(\mathbf{x})||p_0(\mathbf{x})) = E_{\mathbf{t}}[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))] - I(\mathbf{x};\mathbf{t}), \quad (1)$$

where $D(p_1(\mathbf{x})||p_0(\mathbf{x}))$ is KLD, $E_{\mathbf{t}}(\cdot)$ denotes taking expectation according to $\mathbf{t}$, $p_1(\mathbf{x}|\mathbf{t})$ is the conditional PDF of $\mathbf{x}$ conditioned on $\mathbf{t}$ under $\mathcal{H}_1$ and $I(\mathbf{x};\mathbf{t})$ is the MI of $\mathbf{t}$ and $\mathbf{x}$ under $\mathcal{H}_1$. A related result has been used to compute MI in order to obtain the channel capacity per unit cost [10]. The derivation of (1) is straightforward

$$\begin{aligned}\ln\frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} &= \ln\frac{p_1(\mathbf{x}|\mathbf{t})}{p_0(\mathbf{x})} - \ln\frac{p_1(\mathbf{x}|\mathbf{t})}{p_1(\mathbf{x})} \\ &= \ln\frac{p_1(\mathbf{x}|\mathbf{t})}{p_0(\mathbf{x})} - \ln\frac{p_1(\mathbf{x},\mathbf{t})}{p_1(\mathbf{x})\pi(\mathbf{t})}\end{aligned}$$

and taking the expected value with respect to $p_1(\mathbf{x},\mathbf{t})$ produces

$$E_{\mathbf{x},\mathbf{t}}\left[\ln\frac{p_1(\mathbf{x})}{p_0(\mathbf{x})}\right] = E_{\mathbf{t}}E_{\mathbf{x}|\mathbf{t}}\left[\ln\frac{p_1(\mathbf{x}|\mathbf{t})}{p_0(\mathbf{x})}\right] - E_{\mathbf{x},\mathbf{t}}\left[\ln\frac{p_1(\mathbf{x},\mathbf{t})}{p_1(\mathbf{x})\pi(\mathbf{t})}\right] \quad (2)$$

to yield (1). Also, $p_1(\mathbf{x})$ can be written as an averaged conditional PDF by averaging $p_1(\mathbf{x}|\mathbf{t})$ over $\mathbf{t}$, as

$$p_1(\mathbf{x}) = \int_{\mathbf{t}} p(\mathbf{x},\mathbf{t})d\mathbf{t} = \int_{\mathbf{t}} p_1(\mathbf{x}|\mathbf{t})p(\mathbf{t})d\mathbf{t} \quad (3)$$

Thus the term $D(p_1(\mathbf{x})||p_0(\mathbf{x}))$ is the KLD of the *averaged conditional* PDF $p_1(\mathbf{x})$ from the PDF $p_0(\mathbf{x})$.

Furthermore the MI $I(\mathbf{x};\mathbf{t})$ is also an averaged KLD obtained by averaging KLD of the conditional PDF $p_1(\mathbf{x}|\mathbf{t})$ from the unconditional PDF $p_1(\mathbf{x})$, $D(p_1(\mathbf{x}|\mathbf{t})||p_1(\mathbf{x}))$, over all possible signals $\mathbf{t}$

$$\begin{aligned} I(\mathbf{x};\mathbf{t}) &= \int_{\mathbf{t}}\int_{\mathbf{x}} p_1(\mathbf{x},\mathbf{t}) \ln \frac{p_1(\mathbf{x},\mathbf{t})}{p_1(\mathbf{x})p(\mathbf{t})} d\mathbf{x}d\mathbf{t} \\ &= \int_{\mathbf{t}}\int_{\mathbf{x}} p(\mathbf{t})p_1(\mathbf{x}|\mathbf{t}) \ln \frac{p_1(\mathbf{x}|\mathbf{t})}{p_1(\mathbf{x})} d\mathbf{x}d\mathbf{t} \\ &= \int_{\mathbf{t}} p(\mathbf{t}) D(p_1(\mathbf{x}|\mathbf{t})||p_1(\mathbf{x}))d\mathbf{t} \end{aligned} \quad (4)$$

Therefore, all three terms of the decomposition (1) can be interpreted respectively as a special distance measurement in the KLD sense. Alternatively, we can write the relationship as [6]

$$\underbrace{D(p_1(\mathbf{x})||p_0(\mathbf{x}))}_{\text{KLD}} = \underbrace{E_t[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))]}_{\text{SNR}} - \underbrace{I(\mathbf{x};\mathbf{t})}_{\text{MI}}. \quad (5)$$

A simple example is next given to illustrate this important relationship. Assume $\mathbf{t}, \mathbf{w}$ are both independent $N\times 1$ random vectors and have distributions as $\mathbf{t} \sim N(\mathbf{0}, \sigma_t^2\mathbf{I})$ and $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ respectively. Then we have

$$\begin{aligned} \mathbf{x} &\sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \text{ under } \mathcal{H}_0 \\ \mathbf{x} &\sim N\left(\mathbf{0}, (\sigma^2+\sigma_t^2)\mathbf{I}\right) \text{ under } \mathcal{H}_1 \end{aligned}$$

The KLD term is

$$\begin{aligned} D(p_1(\mathbf{x})||p_0(\mathbf{x})) &= \frac{1}{2}\ln\frac{|\sigma^2\mathbf{I}|}{|(\sigma^2+\sigma_t^2)\mathbf{I}|} \\ &\quad + \frac{1}{2}\text{tr}\left[(\sigma^2+\sigma_t^2)\mathbf{I}(\sigma^2\mathbf{I})^{-1} - \mathbf{I}\right] \\ &= \frac{N}{2}\frac{\sigma_t^2}{\sigma^2} - \frac{N}{2}\ln\left(1+\frac{\sigma_t^2}{\sigma^2}\right). \end{aligned} \quad (6)$$

Next, for a given $\mathbf{t}$, the conditional PDF $p_1(\mathbf{x}|\mathbf{t})$ is a Gaussian distribution with mean $\mathbf{t}$ and variance $\sigma^2\mathbf{I}$, so

$$D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x})) = \frac{1}{2}\frac{\mathbf{t}^T\mathbf{t}}{\sigma^2}.$$

Thus, we have

$$\begin{aligned} E_{\mathbf{t}}[D(p_1(\mathbf{x}|\mathbf{t})||p_0(\mathbf{x}))] &= \int_{\mathbf{t}} p(\mathbf{t})\frac{1}{2}\frac{\mathbf{t}^T\mathbf{t}}{\sigma^2}d\mathbf{t} \\ &= \frac{N}{2}\frac{\sigma_t^2}{\sigma^2} \end{aligned}$$

which is indeed a measure of SNR. Lastly, it is easy to show that

$$I(\mathbf{x};\mathbf{t}) = \frac{N}{2}\ln\left(1+\frac{\sigma_t^2}{\sigma^2}\right).$$

Clearly, (5) applies to this simple example. This relationship (5) provides many insights into various problems. For instance, it suggests that MI measures the loss in detection performance between a matched filter, which is based on $\mathbf{t}$ known, and an estimator-correlator, which is based on an average $\mathbf{t}$ [6]. In this paper, however, we focus on using the relationship to justify the meaning of EEF penalty term. This hopefully will further the understanding of the problem of discrimination between normal linear models in [12].

III. Introduction of EEF

Assume that we have two distinct PDFs $p_1(\mathbf{x})$ and $p_0(\mathbf{x})$, and they model the data $\mathbf{x} = [x_0\ x_1\ \cdots\ x_{N-1}]^T$ under a general alternative model hypothesis $\mathcal{H}_1$ and a reference hypothesis $\mathcal{H}_0$. The EEF, denoted as $p(\mathbf{x},\eta)$, is an exponential embedded PDF parameterized by an embedding parameter η, which takes on values $0 \le \eta \le 1$.

$$p(\mathbf{x};\eta) = \frac{p_1^{\eta}(\mathbf{x})p_0^{1-\eta}(\mathbf{x})}{\int p_1^{\eta}(\mathbf{x})p_0^{1-\eta}(\mathbf{x})d\mathbf{x}}. \quad (7)$$

Equivalently, the EEF is expressed as [8]

$$p(\mathbf{x};\eta) = \exp\left[\eta T(\mathbf{x}) - K_0(\eta) + \ln p_0(\mathbf{x})\right]$$

where $T(\mathbf{x}) = \ln\frac{p_1(\mathbf{x})}{p_0(\mathbf{x})}$, $K_0(\eta) = \ln E_0(\exp(\eta T(\mathbf{x})))$, and $E_0(\cdot)$ denotes expectation under $\mathcal{H}_0$. If the PDF $p_1(\mathbf{x}) = p(\mathbf{x};\boldsymbol{\theta})$ has unknown parameters $\boldsymbol{\theta}$, a $p\times 1$ vector and under $\mathcal{H}_0$, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, then upon taking a reduced form and using an asymptotic approximation for the PDF, the EEF reduces to [8]

$$\text{EEF} = \max_{\eta}[\eta \ln\frac{1}{p_{T'}(T'(\mathbf{x});\boldsymbol{\theta}_0)} - K_0(\eta)]$$

where $T'(\mathbf{x}) = \ln\frac{p(\mathbf{x};\hat{\boldsymbol{\theta}})}{p(\mathbf{x};\boldsymbol{\theta}_0)}$ and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the $\boldsymbol{\theta}$.

IV. EEF penalty term-DC level in WGN

In this section we start the discussion of the penalty term of the EEF with a familiar example $\mathbf{x} = A\mathbf{1} + \mathbf{w}$, where A is assumed to be an unknown scalar, $\mathbf{w}$ is white Gaussian noise (WGN) with covariance $\sigma^2\mathbf{I}$, and $\mathbf{1} = [1\ 1\ \cdots\ 1]^T$ is a $N\times 1$ vector. The EEF, termed EEF_d, where the subscript "d" indicates that A is assumed deterministic, is given in [8] as

$$\text{EEF}_\text{d} = \max_{\eta}\left(\eta\frac{N\bar{x}^2}{2\sigma^2} + \frac{1}{2}\ln(1-\eta)\right),$$

where $\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1}x_n$. With $\hat{\eta} = 1 - \frac{\sigma^2}{N\bar{x}^2}$ ($\hat{\eta} = 0$ if $N\bar{x}^2 < \sigma^2$), we have for $0 < \hat{\eta} < 1$

$$\text{EEF}_\text{d} = \frac{1}{2}\left(\frac{N\bar{x}^2}{\sigma^2} - 1\right) - \frac{1}{2}\ln\left(\frac{N\bar{x}^2}{\sigma^2}\right).$$

To verify the relationship between KLD, SNR and MI, we now assume the DC level A is a zero-mean Gaussian random variable with variance $k\frac{\sigma^2}{N}$ instead, and let $k\to\infty$. That is, we assign a vague proper prior to the unknown parameter in an attempt to assigning a non-informative prior. Then, we have

$$\begin{aligned} \mathcal{H}_0 : \mathbf{x} &\sim N(\mathbf{0}, \sigma^2\mathbf{I}) \\ \mathcal{H}_1 : \mathbf{x} &\sim N(\mathbf{0}, \sigma^2\mathbf{I} + k\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T), \end{aligned}$$

and the resultant EEF PDF $p_\eta(\mathbf{x})$ can be shown to be

$$p_\eta(\mathbf{x}) = N(\mathbf{0}, \sigma^2\mathbf{I} + \frac{\eta}{1-\eta}\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T). \quad (8)$$

Proof: $p_\eta(\mathbf{x})$ is an exponential embedding of two zero mean normal distributions PDFs with variance matrices as $\mathbf{C}_0 = \sigma^2\mathbf{I}$ and $\mathbf{C}_1 = \sigma^2\mathbf{I} + k\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T$ respectively. According

to (7), the resultant EEF $p_\eta(\mathbf{x})$ is also a zero mean normal distribution with variance matrix $\mathbf{C}(\eta)$, depending on η, as [8]

$$\begin{aligned}
\mathbf{C}(\eta) &= \left(\eta\mathbf{C}_1^{-1} + (1-\eta)\mathbf{C}_0^{-1}\right)^{-1} \\
&= \left(\eta[\sigma^2\mathbf{I} + k\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T]^{-1} + \frac{1-\eta}{\sigma^2}\mathbf{I}\right)^{-1} \\
&= \left(\frac{\eta}{\sigma^2}\left[\mathbf{I} - \frac{k}{k+1}\frac{1}{N}\mathbf{1}\mathbf{1}^T\right] + \frac{1-\eta}{\sigma^2}\mathbf{I}\right)^{-1} \\
&= \left(\frac{1}{\sigma^2}\left[\mathbf{I} - \frac{\eta k}{k+1}\frac{1}{N}\mathbf{1}\mathbf{1}^T\right]\right)^{-1} \\
&= \sigma^2\left(\mathbf{I} - \frac{-\frac{\eta k}{k+1}}{\frac{-\eta k}{k+1}+1}\frac{1}{N}\mathbf{1}\mathbf{1}^T\right) \\
&\overset{k\to\infty}{\to} \sigma^2\mathbf{I} + \frac{\eta}{1-\eta}\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T
\end{aligned}$$

■

We denote $\mathbf{C}_\eta = \sigma^2\mathbf{I} + \frac{\eta}{1-\eta}\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T$. Alternatively, we consider to assign a prior to A_η,

$$\pi(A_\eta) = N(0, \frac{\eta}{1-\eta}\frac{\sigma^2}{N})$$

for the following model

$$\mathbf{x}_\eta = A_\eta\mathbf{1} + \mathbf{w}$$

Then we have $p_\eta(\mathbf{x}) = p(\mathbf{x}_\eta)$; that is, the two are equivalent PDFs. This shows that EEF method can use vague proper prior and can find an equivalent PDF with a prior on unknown parameter related to the embedding parameter η. This will be proved rigorously in an extended paper. On the other hand, it is generally a bad idea for many other Bayesian model selection methods to use vague proper prior[13].

Then, the EEF for this case, termed EEF_r, where the subscript "r" indicates that A is considered to be the outcome of a random variable, is KLD $D(p_{\hat\eta}(\mathbf{x})||p_0(\mathbf{x}))$. To compute it, we first should find the $\hat\eta$. It is also the value of η that maximizes the following likelihood ratio [8].

$$\begin{aligned}
L_\eta(\mathbf{x}) &= 2\ln\frac{p_\eta(\mathbf{x})}{p_0(\mathbf{x})} \\
&= 2\ln\frac{\frac{1}{\sqrt{(2\pi)^N|\mathbf{C}_\eta|}}\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{C}_\eta^{-1}\mathbf{x})}{\frac{1}{\sqrt{(2\pi)^N|\sigma^2\mathbf{I}|}}\exp(-\frac{1}{2}\mathbf{x}^T(\sigma^2\mathbf{I})^{-1}\mathbf{x})} \\
&= \mathbf{x}^T\left[(\sigma^2\mathbf{I})^{-1} - \mathbf{C}_\eta^{-1}\right]\mathbf{x} - \ln\frac{|\sigma^2\mathbf{I} + \frac{\eta}{1-\eta}\frac{\sigma^2}{N}\mathbf{1}\mathbf{1}^T|}{|\sigma^2\mathbf{I}|} \\
&= \frac{\eta}{N\sigma^2}\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x} - \ln\left|\mathbf{I} + \frac{\eta}{1-\eta}\frac{1}{N}\mathbf{1}\mathbf{1}^T\right| \\
&= \frac{\eta}{N\sigma^2}\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x} - \ln\left(1 + \frac{\eta}{1-\eta}\right) \qquad (9)
\end{aligned}$$

Then the $\hat\eta$ is the value of η for which the derivative is equal to zero and hence, solves the equation

$$\frac{\partial L_\eta(\mathbf{x})}{\partial\eta} = \frac{1}{N\sigma^2}\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x} - \frac{1}{1-\eta} \qquad (10)$$

Incorporating the definition of the embedding parameter $0 \le \eta \le 1$, we have

$$\hat\eta = \begin{cases} 0 & \text{if } \mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x} < N\sigma^2 \\ \frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x} - N\sigma^2}{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}} & \text{otherwise} \end{cases}$$

When $\hat\eta = 0$, the corresponding EEF_r penalty term is zero. We focus on the third case, $0 < \hat\eta < 1$, in the rest of the paper, which is of main interest. The resulting EEF_r is

$$\begin{aligned}
\mathrm{EEF}_r &= D(p_{\hat\eta}||p_0) \\
&= \frac{1}{2}\mathrm{tr}\left[\frac{\hat\eta^2\sigma_A^2}{\sigma^2}\mathbf{1}\mathbf{1}^T\right] - \frac{1}{2}\ln\frac{|\hat\eta^2\sigma_A^2\mathbf{1}\mathbf{1}^T + \sigma^2\mathbf{I}|}{|\sigma^2\mathbf{I}|} \\
&= \frac{1}{2}\frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}}{N\sigma^2} - \frac{1}{2} - \frac{1}{2}\ln\left(\frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}}{N\sigma^2}\right) \\
&= \frac{1}{2}\left(\frac{N\bar x^2}{\sigma^2} - 1\right) - \frac{1}{2}\ln\left(\frac{N\bar x^2}{\sigma^2}\right). \qquad (11)
\end{aligned}$$

This shows that $\mathrm{EEF_r} = \mathrm{EEF_d}$, that is, the resulting EEFs for the two different problems of a deterministic A and a random A are the same. Note that when taking expectation according to $p_{\hat\eta}(\mathbf{x})$, $\hat\eta$ is considered as a constant parameter, not a function of $\mathbf{x}$.

It is easy to prove that the penalty term of $\mathrm{EEF_r}$, $\frac{1}{2}\ln\left(\frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}}{N\sigma^2}\right)$ is indeed $I(\mathbf{x}_{\hat\eta}; A_{\hat\eta})$, the mutual information, since we have

$$\begin{aligned}
I(\mathbf{x}_{\hat\eta}; A_{\hat\eta}) &= E_{A_{\hat\eta}}D(p(\mathbf{x}_{\hat\eta}|A_{\hat\eta})||p(\mathbf{x}_{\hat\eta})) \\
&= \frac{1}{2}\ln\left(\frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x}}{N\sigma^2}\right)
\end{aligned}$$

Strictly speaking, it is an estimated mutual information in that we only have the estimated PDF $p_{\hat\eta}(\mathbf{x})$, or equivalently $p(\mathbf{x}_{\hat\eta})$, instead of the true PDF.

This is a direct result of the equivalency of $p_\eta(\mathbf{x})$ and $p(\mathbf{x}_\eta)$ and the decomposition (5) when applied under the estimated PDF $p_{\hat\eta}(\mathbf{x})$ since the reduced EEF_r is an asymptotic KLD $D(p_{\hat\eta}||p_0)$. A modified version of decomposition (5) can be expressed as follows

$$\mathrm{EEF}_r = \widehat{\mathrm{SNR}} - \widehat{\mathrm{MI}}$$

where $\widehat{\mathrm{SNR}}, \widehat{\mathrm{MI}}$ are the estimated SNR and estimated MI, respectively.

V. EEF PENALTY TERM OF LINEAR MODEL

We now generalize the previous results to show that the EEF penalty term of model order selection for the linear model is the estimated MI. The linear model is an important one in practice and so a detailed analysis of this result is warranted. The linear model is $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ where $\mathbf{H}$ is $N \times p$, $\boldsymbol{\theta}$ is a $p \times 1$ vector and $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Different models have different orders p and observation matrices $\mathbf{H}$. The model order selection problem is to decide the value of p to best model the data. It can be shown that assuming $\boldsymbol{\theta}$ is a deterministic unknown parameter yields the same EEF as assuming it is a random vector with a given prior PDF [8],[9]. We assume the latter by assigning to the unknown parameter

$\boldsymbol{\theta}$ the prior PDF $N(\mathbf{0}, \frac{\xi^2}{p}\sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$. When ξ^2 is assumed unknown, the EEF is proved to be equivalent to the model structure determination (MSD) [9]. If we reparameterize ξ^2 by letting

$$\frac{\xi^2}{p} = \frac{\eta}{1-\eta},$$

then a one-to-one transformation from ξ^2 to η $(0 < \eta < 1)$ is effected and finding $\hat{\eta}$ is equivalent to finding $\hat{\xi^2}$. With this setup, we have

$$\mathbf{x} \sim N(\mathbf{0}, \sigma^2\mathbf{I} + \frac{\xi^2}{p}\sigma^2\mathbf{P_H}) \text{ under } p_\eta(\mathbf{x})$$

where $\mathbf{P_H} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$. It is shown in [9] that the EEF for model $\mathcal{M}_p$, i.e., with p unknown parameters, is

$$\mathrm{EEF}_r(p) = \max_{\frac{\xi^2}{p}} \left[\frac{1}{2\sigma^2}\frac{\frac{\xi^2}{p}}{1+\frac{\xi^2}{p}}\mathbf{x}^T\mathbf{P_H}\mathbf{x} - \frac{p}{2}\ln(1+\frac{\xi^2}{p})\right].$$

The estimate $\frac{\hat{\xi^2}}{p}$, which maximizes $\mathrm{EEF}_r(p)$ is

$$\frac{\hat{\xi^2}}{p} = \frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{p\sigma^2} - 1$$

and hence, the maximized EEF is

$$\mathrm{EEF}_r(p) = \frac{1}{2}\left(\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2} - p\right) - \frac{p}{2}\ln\frac{\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2}}{p}. \qquad (12)$$

Since $\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2}$ obeys a χ^2_p distribution under the null hypothesis [7],[8], the term $\frac{1}{2}\left(\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2} - p\right)$ subtracts out the mean p under $\mathcal{H}_0$, thereby producing $\widehat{\mathrm{SNR}}$. The term $\frac{p}{2}\ln\frac{\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2}}{p}$ is the estimated $\widehat{\mathrm{MI}}$ as shown next. First, we have the following

$$\begin{aligned}
&D(p_\eta(\mathbf{x}|\boldsymbol{\theta})||p_\eta(\mathbf{x})) \\
&= \frac{1}{2}\ln\frac{|\sigma^2\mathbf{I} + \frac{\xi^2}{p}\sigma^2\mathbf{P_H}|}{|\sigma^2\mathbf{I}|} + \frac{1}{2}\mathrm{tr}\left(\sigma^2(\sigma^2\mathbf{I} + \frac{\xi^2}{p}\sigma^2\mathbf{P_H})^{-1} - \mathbf{I}\right) \\
&\quad + \frac{1}{2}(\mathbf{H}\boldsymbol{\theta})^T(\sigma^2\mathbf{I} + \frac{\xi^2}{p}\sigma^2\mathbf{P_H})^{-1}\mathbf{H}\boldsymbol{\theta} \\
&= \frac{1}{2}\ln\frac{|\sigma^2\mathbf{I} + \frac{\xi^2}{p}\sigma^2\mathbf{P_H}|}{|\sigma^2\mathbf{I}|} + \frac{1}{2}\mathrm{tr}\left(-\frac{\frac{\xi^2}{p}}{\frac{\xi^2}{p}+1}\mathbf{P_H}\right) \\
&\quad + \frac{1}{2\sigma^2}(\mathbf{H}\boldsymbol{\theta})^T(\mathbf{I} - \frac{\frac{\xi^2}{p}}{\frac{\xi^2}{p}+1}\mathbf{P_H})\mathbf{H}\boldsymbol{\theta}.
\end{aligned}$$

Then the computation of the estimated MI between $\mathbf{x}$ and $\boldsymbol{\theta}$ follows.

$$\begin{aligned}
I_{\hat{\eta}}(\mathbf{x};\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}D(p_{\hat{\eta}}(\mathbf{x}|\boldsymbol{\theta})||p_{\hat{\eta}}(\mathbf{x})) \\
&= \frac{1}{2}\ln\frac{|\sigma^2\mathbf{I} + \frac{\hat{\xi^2}}{p}\sigma^2\mathbf{P_H}|}{|\sigma^2\mathbf{I}|} + \frac{1}{2}\mathrm{tr}\left(-\frac{\frac{\hat{\xi^2}}{p}}{\frac{\hat{\xi^2}}{p}+1}\mathbf{P_H}\right) \\
&\quad + E_{\boldsymbol{\theta}}\left[\frac{1}{2\sigma^2}\frac{1}{\frac{\hat{\xi^2}}{p}+1}(\mathbf{H}\boldsymbol{\theta})^T\mathbf{H}\boldsymbol{\theta}\right] \\
&= \frac{1}{2}\ln\frac{|\sigma^2\mathbf{I} + \frac{\hat{\xi^2}}{p}\sigma^2\mathbf{P_H}|}{|\sigma^2\mathbf{I}|} = \frac{p}{2}\ln\frac{\frac{\mathbf{x}^T\mathbf{P_H}\mathbf{x}}{\sigma^2}}{p} \qquad (13)
\end{aligned}$$

where we have applied

$$E_{\boldsymbol{\theta}}\left[\frac{1}{2\sigma^2}\frac{1}{\frac{\hat{\xi^2}}{p}+1}(\mathbf{H}\boldsymbol{\theta})^T\mathbf{H}\boldsymbol{\theta}\right] = \frac{1}{2\sigma^2}\frac{1}{\frac{\hat{\xi^2}}{p}+1}\mathrm{tr}(\frac{\hat{\xi^2}}{p}\sigma^2\mathbf{P_H}),$$

$(\mathbf{I} + \frac{\hat{\xi^2}}{p}\mathbf{P_H})^{-1} = \mathbf{I} - \frac{\frac{\hat{\xi^2}}{p}}{\frac{\hat{\xi^2}}{p}+1}\mathbf{P_H}$ and $\left|\mathbf{I} + \frac{\hat{\xi^2}}{p}\mathbf{P_H}\right| = \left(1 + \frac{\hat{\xi^2}}{p}\right)^p$. Thus, (13) proves that the EEF penalty term for the linear model in (12) is indeed the estimated MI. This is intuitively appealing in the sense that the model order selection rule should not take into account the information contributed by the distributional knowledge of the unknown parameters, which increases with its dimension [12]. As a special case, when $\mathbf{H} = \mathbf{1}$ and $\boldsymbol{\theta} = A$ then this example reduces to the DC level in WGN example for which $p = 1$. Thus, the estimated MI term in (13) reduces to $\frac{1}{2}\ln\left(\frac{\mathbf{x}^T\mathbf{1}\mathbf{1}^T}{N\sigma^2}\right)$, which is the estimated MI in (11).

VI. Conclusions

To summarize, we show that EEF method can employ vague proper priors in model order selection. The resultant penalty term can be viewed as an estimated mutual information between model unknown parameters and the received data from Bayesian viewpoints. Intuitively, the MI measures how much information of the data is contributed by the parameter $\boldsymbol{\theta}$. The EEF model order selection rule therefore subtracts it out so that the comparison among different models tends to be more fair. Future work will discuss the relationship of the estimated mutual information and other concepts in model selection such as model complexity.

References

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

[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," *Ann. Statist.,* vol.6, no.2, pp.461–464,1978.

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

[6] Z. Zhu, S. Kay, and R.S. Raghavan,"Information-theoretical optimal radar waveform design," *submitted to the IEEE Signal Processing Letters*.

[7] S. Kay, *Fundamentals of Statistical Signal Processing: Detection*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[8] S. Kay. "Exponentially embedded families-New approaches to model order estimation", *IEEE Trans. on Aerospace and Electronic Systems*, vol.41, no.1, pp.333–344, Jan. 2005.

[9] S. Kay and Q. Ding, "Model estimation and classification via model structure determination", *IEEE Trans. on Signal Processing*,vol. 61, no.10, pp 2588-2597, 2013

[10] S. Verdu, "On channel capacity per unit cost," *IEEE Trans. Inform. Theory,* vol. 36, pp. 1019–1030, Sept. 1990.

[11] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Processing,* vol.46, no.10, pp. 2726–2735, Oct. 1998.

[12] L.R. Pericchi, "An alternative to the standard Bayesian procedure for discrimination between normal linear models", *Biometrika*, vol.71, no.3, pp.575–586, Dec. 1984.

[13] J. Berger and L. Pericchi, "Objective Bayesian methods for model selection: Introduction and comparison," *in Model Selection, vol. 38 of IMS Lecture Notes–Monograph Series, (ed. P.Lahiri)*, pp.135-193, Institute of Mathematical Statistics, 2001.