EMPIRICAL BAYES LINEAR REGRESSION WITH UNKNOWN MODEL ORDER

Yngve Selén

Dept. of Information Technology
Uppsala University
P.O. Box 337, 751 05 Uppsala, Sweden.
Email: yngve.selen@it.uu.se.

Erik G. Larsson

School of EE, Communication Theory
Royal Institute of Technology (KTH)
Osquldas väg 10, 100 44 Stockholm, Sweden
Email: erik.larsson@ee.kth.se.

ABSTRACT

We study the maximum *a posteriori* probability model order selection algorithm for linear regression models, assuming Gaussian distributed noise and coefficient vectors. For the same data model, we also derive the minimum mean-square error coefficient vector estimate. The approaches are denoted BOSS (Bayesian Order Selection Strategy) and BPM (Bayesian Parameter estimation Method), respectively. Both BOSS and BPM require *a priori* knowledge on the distribution of the coefficients. However, under the assumption that the coefficient variance profile is smooth, we derive "empirical Bayesian" versions of our algorithms, which require little or no information from the user. We show in numerical examples that the estimators can outperform several classical methods, including the well-known AIC and BIC for order selection.

***Index Terms*—** Linear systems, Bayes procedures, modeling, least mean square methods, parameter estimation

1. INTRODUCTION

1.1. Problem Formulation

Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{h} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^N$ is the vector of observed data, $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_n] \in \mathbb{R}^{N\times n}$ is a known matrix of n regressors $\{\boldsymbol{x}_j\}_{j=1}^n$, $\boldsymbol{h} = [h_1 \cdots h_n]^T \in \mathbb{R}^n$ is the unknown coefficient vector and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ is a length N vector of zero-mean Gaussian white noise with variance σ^2. We call (1) the *full model* and assume that the data are generated by a model of the form

$$\mathcal{M}_k : \boldsymbol{y} = \boldsymbol{X}_k\boldsymbol{h}_k + \boldsymbol{\epsilon} \tag{2}$$

where $n_{\min} \le k \le n$, $\boldsymbol{X}_k = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_k]$ (i.e., $\boldsymbol{X}_k$ consists of the first k columns of $\boldsymbol{X}$), and $\boldsymbol{h}_k = [h_1 \cdots h_k]^T$. Furthermore, we make the assumption that the coefficients h_j are zero-mean independent Gaussian random variables, $h_j \sim \mathcal{N}(0, \gamma_j^2)$. In other words, $\boldsymbol{h}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_k)$ where $\boldsymbol{\Gamma}_k = \mathrm{diag}[\gamma_1^2 \cdots \gamma_k^2]$. The model order k is assumed to be unknown.

We consider the following two classical interrelated problems:

1. The model order selection problem: to correctly detect the order k, given $\boldsymbol{X}$ and $\boldsymbol{y}$.
2. The parameter estimation problem: to estimate $\boldsymbol{h}$ as accurately as possible, assuming the order k is unknown.

This work was supported in part by the Swedish Science Council (VR).

1.2. Related Work

Bayesian solutions to the above two problems, under the Gaussianity assumption on the coefficients and the noise, are available in the literature. In, e.g., [1], the maximum likelihood (ML) model order selection algorithm for the current model was derived, although not numerically evaluated. In [2], the minimum mean square error (MMSE) estimate of the frequency function was derived. Within the same framework, it is easy to derive the MMSE estimate of $\boldsymbol{h}$. In [3, 4] simple derivations of the maximum *a posteriori* (MAP) model order selection algorithm and the MMSE estimate of $\boldsymbol{h}$ were presented. These derivations will be the basis of the empirical Bayes method that we propose, and will be summarized in Sections 2 and 3.

1.3. Contribution of This Work

In the references in the above subsection it is generally assumed that the noise variance σ^2 and the coefficient variances $\{\gamma_j^2\}_{j=1}^n$ are *known*. This assumption is hardly realistic in applications. The goal of the present article is to present methods which *do not* require knowledge of σ^2 and $\{\gamma_j^2\}_{j=1}^n$. To this end we take an empirical Bayes approach: we estimate σ^2, $\{\gamma_j^2\}_{j=1}^n$ from the data and then use the resulting estimates as if they were the true values. See Section 4.

2. OPTIMAL MODEL ORDER SELECTION

Here we review the maximum *a posteriori* (MAP) probability model order selection algorithm for the problem posed in Section 1, assuming known σ^2, $\{\gamma_j^2\}_{j=1}^n$. Note that this model selection rule has been derived previously in [1]. In the remainder of the paper we will denote this specific model selection algorithm BOSS (Bayesian Order Selection Strategy).

Using Bayes' Theorem we obtain an expression for the model posterior probabilities:

$$P(\mathcal{M}_k|\boldsymbol{y}) = P(\mathcal{M}_k)\frac{p(\boldsymbol{y}|\mathcal{M}_k)}{p(\boldsymbol{y})}.$$

Since $p(\boldsymbol{y})$ is independent of the model $\mathcal{M}_k$, the model order which gives the highest posterior probability model is

$$\mathrm{MAP}: \quad \hat{k} \;=\; \arg\max_{k=n_{\min},\ldots,n} P(\mathcal{M}_k)p(\boldsymbol{y}|\mathcal{M}_k). \tag{3}$$

If nothing is known about the model prior probabilities $P(\mathcal{M}_k)$, we will assume that they are equal and, of course, that they sum up to one:

$$P(\mathcal{M}_k) = \frac{1}{n - n_{\min} + 1}, \qquad k = n_{\min}, \ldots, n \tag{4}$$

(this is common practice [5]). Furthermore, under the assumption that $\mathcal{M}_k$ is the data generating model we have

$$\boldsymbol{y}|\mathcal{M}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}_k)$$

where $\boldsymbol{Q}_k = \boldsymbol{X}_k\boldsymbol{\Gamma}_k\boldsymbol{X}_k^T + \sigma^2\boldsymbol{I}$, $\boldsymbol{\Gamma}_k = \mathrm{diag}[\gamma_1^2 \cdots \gamma_k^2]$. So,

$$p(\boldsymbol{y}|\mathcal{M}_k) = \frac{1}{\sqrt{2\pi}^N}\frac{1}{|\boldsymbol{Q}_k|^{1/2}}\exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{Q}_k^{-1}\boldsymbol{y}\right). \quad (5)$$

We obtain BOSS by using (4) and (5) to compute (3). One can show (see, e.g., the discussion around Equation (37) in [6]) that the order with the highest *a posteriori* probability is also the most likely to be the correct order.

Note that (using the identity $|\boldsymbol{I} + \boldsymbol{AB}| = |\boldsymbol{I} + \boldsymbol{BA}|$)

$$|\boldsymbol{Q}_k| = \left|\boldsymbol{X}_k\boldsymbol{\Gamma}_k\boldsymbol{X}_k^T + \sigma^2\boldsymbol{I}\right| = \sigma^{2N}\left|\frac{\boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{\Gamma}_k}{\sigma^2} + \boldsymbol{I}\right| \quad (6)$$

and (using the matrix inversion lemma)

$$\begin{aligned}\boldsymbol{Q}_k^{-1} &= \left(\boldsymbol{X}_k\boldsymbol{\Gamma}_k\boldsymbol{X}_k^T + \sigma^2\boldsymbol{I}\right)^{-1} \\ &= \frac{1}{\sigma^2}\boldsymbol{I} - \frac{1}{\sigma^4}\boldsymbol{X}_k\left(\boldsymbol{\Gamma}_k^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}_k^T\boldsymbol{X}_k\right)^{-1}\boldsymbol{X}_k^T.\end{aligned} \quad (7)$$

The expressions in (6) and (7) can be used to boost the computational efficiency: $\boldsymbol{X}_k^T\boldsymbol{X}_k$ is much faster to evaluate than $\boldsymbol{X}_k\boldsymbol{X}_k^T$ if $N \gg n$ (one should also exploit the fact that $\boldsymbol{\Gamma}_k$ is diagonal). The computational complexity can be further reduced by noting that

$$\boldsymbol{X}_k^T\boldsymbol{X}_k = \begin{bmatrix}\boldsymbol{X}_{k-1}^T \\ \boldsymbol{x}_k^T\end{bmatrix}[\boldsymbol{X}_{k-1}\ \boldsymbol{x}_k] = \begin{bmatrix}\boldsymbol{X}_{k-1}^T\boldsymbol{X}_{k-1} & (\boldsymbol{x}_k^T\boldsymbol{X}_{k-1})^T \\ \boldsymbol{x}_k^T\boldsymbol{X}_{k-1} & \boldsymbol{x}_k^T\boldsymbol{x}_k\end{bmatrix}$$

and that $\boldsymbol{X}_{k-1}^T\boldsymbol{X}_{k-1}$ is evaluated when computing $p(\boldsymbol{y}|\mathcal{M}_{k-1})$. Also, a well-known lemma on the inverse of partitioned matrices (see, e.g., Lemma A.2 in [7]) can be used to compute (7) iteratively. Define $\boldsymbol{Z}_k = \left(\boldsymbol{\Gamma}_k^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}_k^T\boldsymbol{X}_k\right)$ (see (7)). Then

$$\begin{aligned}\boldsymbol{Z}_k^{-1} &= \begin{bmatrix}\boldsymbol{Z}_{k-1}^{-1} & \boldsymbol{0}_{k-1} \\ \boldsymbol{0}_{k-1}^T & 0\end{bmatrix} \\ &+ \frac{\begin{bmatrix}-\boldsymbol{Z}_{k-1}^{-1}\frac{(\boldsymbol{x}_k^T\boldsymbol{X}_{k-1})^T}{\sigma^2} \\ 1\end{bmatrix}\begin{bmatrix}-\frac{\boldsymbol{x}_k^T\boldsymbol{X}_{k-1}}{\sigma^2}\boldsymbol{Z}_{k-1}^{-1} & 1\end{bmatrix}}{\frac{\boldsymbol{x}_k^T\boldsymbol{x}_k}{\sigma^2} + \frac{1}{\gamma_k^2} - \frac{1}{\sigma^4}\boldsymbol{x}_k^T\boldsymbol{X}_{k-1}\boldsymbol{Z}_{k-1}^{-1}(\boldsymbol{x}_k^T\boldsymbol{X}_{k-1})^T},\end{aligned}$$

so the matrix inversion in (7) needs only be computed directly for $k = n_{\min}$.

3. THE MMSE ESTIMATE OF $\boldsymbol{h}$

In this section we describe what we will denote the BPM (Bayesian Parameter estimation Method) by deriving the MMSE estimate of $\boldsymbol{h}$ under the assumption that one of the models $\{\mathcal{M}_k\}_{k=n_{\min}}^n$ in (2) generated the data. The MMSE estimate equals the conditional mean:

$$\hat{\boldsymbol{h}}_{\text{MMSE}} = E[\boldsymbol{h}|\boldsymbol{y}] = \sum_{k=n_{\min}}^{n} P(\mathcal{M}_k|\boldsymbol{y})E[\boldsymbol{h}|\boldsymbol{y},\mathcal{M}_k]. \quad (8)$$

Note that

$$P(\mathcal{M}_k|\boldsymbol{y}) = P(\mathcal{M}_k)\frac{p(\boldsymbol{y}|\mathcal{M}_k)}{p(\boldsymbol{y})} = \frac{P(\mathcal{M}_k)p(\boldsymbol{y}|\mathcal{M}_k)}{\sum_{j=n_{\min}}^{n} P(\mathcal{M}_j)p(\boldsymbol{y}|\mathcal{M}_j)}. \quad (9)$$

By inserting (9) in (8) we get

$$\text{MMSE}: \quad \hat{\boldsymbol{h}}_{\text{MMSE}} = \frac{\sum_{k=n_{\min}}^{n} P(\mathcal{M}_k)p(\boldsymbol{y}|\mathcal{M}_k)E[\boldsymbol{h}|\boldsymbol{y},\mathcal{M}_k]}{\sum_{k=n_{\min}}^{n} P(\mathcal{M}_k)p(\boldsymbol{y}|\mathcal{M}_k)}. \quad (10)$$

We can use (4) and (5) in (10). What remains to evaluate (10) and get the BPM is then to compute $E[\boldsymbol{h}|\boldsymbol{y},\mathcal{M}_k]$.

Clearly, assuming that the model $\mathcal{M}_k$ generated the data, $h_j = 0$ for $j > k$, so it is sufficient to find $E[\boldsymbol{h}_k|\boldsymbol{y},\mathcal{M}_k]$. Under $\mathcal{M}_k$, $\boldsymbol{h}_k$ and $\boldsymbol{y}$ are jointly Gaussian:

$$\begin{bmatrix}\boldsymbol{y} \\ \boldsymbol{h}_k\end{bmatrix}|\mathcal{M}_k \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix}\boldsymbol{Q}_k & \boldsymbol{X}_k\boldsymbol{\Gamma}_k^T \\ \boldsymbol{\Gamma}_k\boldsymbol{X}_k^T & \boldsymbol{\Gamma}_k\end{bmatrix}\right).$$

Applying a standard result (Lemma B.17 in [7], for example), the conditional mean evaluates to

$$E[\boldsymbol{h}_k|\boldsymbol{y},\mathcal{M}_k] = \boldsymbol{\Gamma}_k\boldsymbol{X}_k^T\boldsymbol{Q}_k^{-1}\boldsymbol{y}. \quad (11)$$

We now obtain the BPM by inserting (4), (5) and (11) into (10).

4. NEW EMPIRICAL BAYESIAN ESTIMATORS

The weakness of BPM and BOSS is that they require knowledge of $\{\gamma_k^2\}_{k=1}^n$, σ^2. In practice, these variances are likely to be unknown. This problem can be dealt with in different ways.

One possibility is to assign hierarchical distributions to the unknown variances and set the resulting hyperparameters to some values (e.g., assume that $\sigma^2 \sim \mathcal{D}(a)$ where $\mathcal{D}$ denotes some distribution and a is a hyperparameter). The resulting expressions, corresponding to (3) and (10), can typically not be solved analytically. Thus one has to resort to numerical approaches, such as Markov Chain Monte-Carlo (MCMC) methods. For a related example, see [8]. This type of numerical approach is generally computationally intensive. Furthermore, it is not clear at what hierarchical level one should stop the hyperparameterization or what values should be assigned to the hyperparameters at the final level: In the above example, should a be set to a specific value, or should we assign yet a distribution for a?

Instead, we propose to use an empirical Bayes approach. In this approach the unknown prior (hyper)parameters are estimated from the available data. The so-obtained estimates of $\{\gamma_k^2\}_{k=1}^n$, σ^2 can then be used for inference, i.e. be inserted into (3) and (10) as if they were the true values. This voids the optimality of BOSS and BPM, as the variances $\{\gamma_k^2\}_{k=1}^n$, σ^2 are *a priori* parameters, which should not depend on the data. Nevertheless, estimation of $\{\gamma_k^2\}_{k=1}^n$, σ^2 appears to be an attractive, pragmatic way of handling the situation when these parameters are completely unknown.

We will base our estimates of $\{\gamma_k^2\}_{k=1}^n$, σ^2 on the least squares (LS) estimator,

$$\hat{\boldsymbol{h}}_{\text{LS},k} = (\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^T\boldsymbol{y}, \quad (12)$$

of the full model (1) (i.e., $k := n$ in (12)). Note that this requires $N \geq n$ and that $\boldsymbol{X}$ is full rank, so these conditions are necessary for our empirical Bayes methods to work.

The estimation of σ^2 is straightforward: an unbiased, consistent estimate of σ^2 can be obtained by taking [4]

$$\hat{\sigma}^2 = \frac{1}{N-n}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{h}}_{\text{LS},n}\|^2. \quad (13)$$

The estimation of $\{\gamma_k^2\}_{k=1}^n$ is more challenging. First, we have only a single realization of $\boldsymbol{y}$. This also means that we have only one single realization of $\boldsymbol{h}$ to use for estimation of $\{\gamma_k^2\}_{k=1}^n$. In other words, only one sample h_k is indirectly available (h_k is unknown, and has to be estimated from $\boldsymbol{y}$) for the estimation of the variance γ_k^2. It is obvious that some *a priori* information is necessary to regularize this problem. Second, if the true order is k, the $\gamma_j^2, j > k$ can not be estimated at all, as there is no data available ($h_j = 0$ when $j > k$). If this problem were disregarded we would likely end up with severely underestimated values of $\hat{\gamma}_j^2$ for $j > k$, which would give a high risk of overestimation of the model order.

Because of the above, it is likely that any estimates of $\{\gamma_k^2\}_{k=1}^n$ will deviate rather much from their true values. Fortunately, BOSS

and BPM appear to be relatively insensitive to mismatching values of $\{\gamma_k^2\}_{k=1}^n$ (see, e.g., [4] and the numerical examples herein).

To cope with the first of the above mentioned problems it is necessary to impose some assumptions on $\{\gamma_k^2\}_{k=1}^n$. We will make the (in our opinion, reasonable) assumption that $\{\gamma_k^2\}_{k=1}^n$ is a *smooth* sequence, i.e. $\gamma_k^2 - \gamma_{k-1}^2$ does not vary a lot with k. In the next subsections we describe two ways of exploiting this assumption. To cope with the second problem above, our solution is to impose a threshold b, such that always $\hat{\gamma}_k^2 \geq b > 0$. In this manner, we need to define the degree of smoothness and the threshold b. We will select the threshold b as a fraction of the largest squared magnitude element in the full model LS estimate ((12) with $k := n$), i.e.

$$b := b_f \max\{|\hat{h}_{\text{LS},1}|^2, |\hat{h}_{\text{LS},2}|^2, \ldots, |\hat{h}_{\text{LS},n}|^2\}. \quad (14)$$

As a rule of thumb, we recommend $b_f = 0.1$. In the following two subsections we describe ways of exploiting the "smoothness constraint". Our proposed approaches require little or no *a priori* information.

4.1. Estimation of $\{\gamma_k^2\}_{k=1}^n$ by Parameterization

In this approach we assume that $\{\gamma_k^2\}_{k=1}^n$ can be expressed as a linear combination of a relatively small number of known basis vectors, as follows:

$$\boldsymbol{\gamma} = \boldsymbol{\Psi}\boldsymbol{\alpha} \quad (15)$$

where $\boldsymbol{\gamma} = [\gamma_1^2 \cdots \gamma_n^2]^T$, $\boldsymbol{\Psi}$ is a known $n \times r$ matrix and $\boldsymbol{\alpha}$ is an unknown length r vector. The columns of $\boldsymbol{\Psi}$ should be a basis for "likely" variance profiles for $\{\gamma_k^2\}_{k=1}^n$ and generally r should be small compared to n. (In a communication application, e.g., it might be known that $\{\gamma_k^2\}_{k=1}^n$ is exponentially decaying [9], and $\boldsymbol{\Psi}$ could then contain one or a few columns with different degrees of exponential decay.) If nothing at all is known about $\{\gamma_k^2\}_{k=1}^n$, then $\boldsymbol{\Psi}$ can be constructed from the first r basis functions in a discrete Fourier series expansion of $\boldsymbol{\gamma}$ (in this case, r is the user parameter):

$$\begin{aligned} [\boldsymbol{\Psi}]_{j,1} &= 1 \\ [\boldsymbol{\Psi}]_{j,k} &= \begin{cases} \cos\left(\frac{k}{2}(j-1)\frac{2\pi}{n}\right) & k \text{ even} \\ \sin\left(\frac{k-1}{2}(j-1)\frac{2\pi}{n}\right) & k \text{ odd and } \geq 3 \end{cases} \end{aligned} \quad (16)$$

where $j = 1, \ldots, n$, $k = 1, \ldots, r$ and r is odd. Using a truncated Fourier series as basis may seem *ad hoc* at first glance, but we find it attractive for many reasons: the basis functions are orthogonal and the resulting variance profile is smooth. Additionally, by choosing the number of basis functions r, we can directly influence the amount of variation in the variance profile $\boldsymbol{\gamma}$.

Now, the problem is reduced to that of estimating $\boldsymbol{\alpha}$, i.e. the dimensionality of the problem is reduced from n to r. Note that

$$E[|h_j|^2] = \sum_{k=n_{\min}}^{n} P(\mathcal{M}_k) E[|h_j|^2 | \mathcal{M}_k] = q_j \gamma_j^2$$

where $q_j = \sum_{k=\max(j,n_{\min})}^{n} P(\mathcal{M}_k)$, because under $\mathcal{M}_k$, $h_j = 0$ for $j > k$, and h_j has variance γ_j^2 if $j \leq k$. By replacing $E[|h_j|^2]$ in the above equation by the squared norm of the corresponding element of the full model LS estimate ((12) with $k := n$), $|\hat{h}_{\text{LS},j}|^2$, and using (15) we can estimate $\boldsymbol{\alpha}$ from the following constrained LS expression:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\Psi}\boldsymbol{\alpha} \geq \mathbf{0}} \left\| \begin{bmatrix} |\hat{h}_{\text{LS},1}|^2 \\ \vdots \\ |\hat{h}_{\text{LS},n}|^2 \end{bmatrix} - \begin{bmatrix} q_1\Psi_{1,1} & \cdots & q_1\Psi_{1,r} \\ \vdots & & \vdots \\ q_n\Psi_{n,1} & \cdots & q_n\Psi_{n,r} \end{bmatrix} \boldsymbol{\alpha} \right\|^2. \quad (17)$$

The constraint in (17) guarantees that all $\hat{\gamma}_j^2 \geq 0$. The above expression can be easily and efficiently solved by quadratic programming (use, e.g., `quadprog` in Matlab).

One might inquire as to why the threshold b is not used already in (17) (and (18) below). The reason is that the thresholds $\boldsymbol{\gamma} \geq \mathbf{0}$ and $\hat{\gamma}_k^2 \geq b$ serve different purposes: $\boldsymbol{\gamma} \geq \mathbf{0}$ used in (17), (18) takes care of the physical requirements on the variances, whereas the threshold b is a way to cope with the lack of data for estimation of $\{\hat{\gamma}_k^2\}_{k=1}^n$. We found that this way of imposing thresholds on $\{\hat{\gamma}_k^2\}_{k=1}^n$ gave the best numerical performance.

4.2. Estimation of $\{\gamma_k^2\}_{k=1}^n$ by Penalization

The problem we would like to solve—that of estimating $\{\gamma_k^2\}_{k=1}^n$ using the LS estimate $\hat{\boldsymbol{h}}_{\text{LS},n}$ from a single data realization—is an ill-conditioned problem. Tikhonov regularization [10] is a commonly used method for solving ill-posed problems. The regularization consists of an additive penalty term which depends on the parameter of interest. This penalty can be designed to, e.g., shrink the estimate towards zero, or limit its variability. Using the *a priori* assumption that $\{\gamma_k^2\}_{k=1}^n$ is a smooth sequence, we impose a penalty on its second order difference and estimate it from

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma} \geq \mathbf{0}} \|\boldsymbol{\gamma} - [|\hat{h}_{\text{LS},1}|^2 \cdots |\hat{h}_{\text{LS},n}|^2]^T\|^2 + \lambda \|\boldsymbol{L}\boldsymbol{\gamma}\|^2. \quad (18)$$

Here, $\boldsymbol{L} = \begin{bmatrix} 1 & -2 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(n-2)\times n}$ is the second order difference matrix and λ is a user parameter which determines the amount of smoothing. Equation (18) can be efficiently solved using quadratic programming. λ should be set to a low value if $\{\gamma_k^2\}_{k=1}^n$ is believed to have large variations, and a high value if $\{\gamma_k^2\}_{k=1}^n$ is believed to be very smooth (the extreme case $\lambda = \infty$ forces $\gamma_k^2 - \gamma_{k-1}^2$ to be constant with respect to k). Alternatively, λ can be selected in a fully automatic manner (i.e., without any user parameters) by using, e.g., generalized cross-validation (GCV). An efficient implementation of the GCV for the problem under consideration is given in [11].

5. NUMERICAL EXAMPLES

We evaluate the performances of the methods by means of Monte-Carlo simulations. The performance of BOSS (3) is measured in terms of the percentage of correctly selected orders. For the evaluation of BPM (10) we use the empirical MSE of the coefficient estimates: $M^{-1}\sum_{m=1}^{M} \|\hat{\boldsymbol{h}}^{(m)} - \boldsymbol{h}^{(m)}\|^2$, where $\hat{\boldsymbol{h}}^{(m)}$ and $\boldsymbol{h}^{(m)}$ denote the estimated and true coefficient vectors (zero-padded if necessary) for realization number m. M is the total number of Monte-Carlo runs and we choose $M = 10^5$.

For each Monte-Carlo trial we generate data from a model $\mathcal{M}_k$ where the order k is chosen uniformly at random between $n_{\min} = 1$ and $n = 30$ (these limits are supplied to the estimators). We set $N = 50$. Our regressor matrix $\boldsymbol{X}$ is composed of i.i.d. $\mathcal{N}(0,1)$ elements. The true variance profiles, $\{\gamma_k^2\}_{k=1}^n$ are constructed from (15) where the first column of $\boldsymbol{\Psi}$ consists of only ones $[1 \cdots 1]^T$, the second is exponentially decaying $[\nu_1 e^{-0.4} \cdots \nu_1 e^{-0.4n}]^T$ and the third is exponentially increasing $[\nu_2 e^{0.4} \cdots \nu_2 e^{0.4n}]^T$. The normalization factors ν_1 and ν_2 are set such that all columns in $\boldsymbol{\Psi}$ have a 2-norm equal to n. The vector $\boldsymbol{\alpha}$ has independent squared $\mathcal{N}(0,1)$ elements (i.e., each element is generated from a $\chi^2(1)$-distribution). Finally, the so generated $\{\gamma_k^2\}_{k=1}^n$ are normalized such that $\|\boldsymbol{\gamma}\|^2 = 1$. Naturally, these choices are somewhat arbitrary (as any specific numerical example has to be), but having compared with other examples we believe this to show a fair comparison of the considered methods.

We consider the following methods: (a) The well-known corrected (for short data sequences) information criterion by Akaike (AICc) [12] for order selection (for parameter estimation we use LS

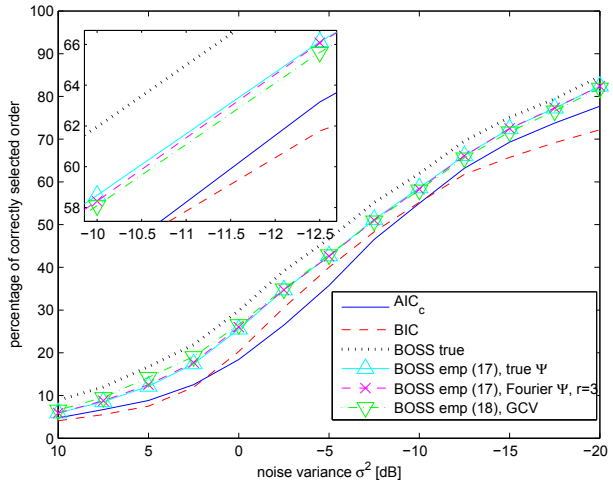

Fig. 1. Model order selection performance: Percentages of correctly selected order. (The small plot is a closeup.)

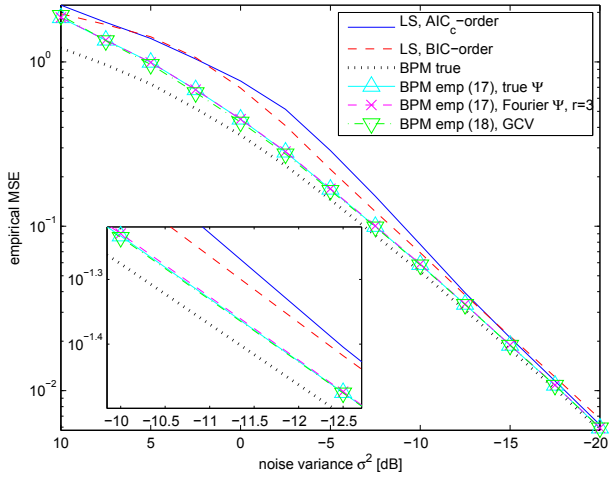

Fig. 2. Coefficient estimation performance: Empirical MSEs. (The small plot is a closeup.)

(12) with the AICc order); (b) The well-known Bayesian information criterion (BIC) [13] for order selection (for parameter estimation we use LS (12) with the BIC order); (c) BOSS/BPM ((3)/(10)) with knowledge of the true $\{\gamma_k^2\}_{k=1}^n$, σ^2; (d) Empirical BOSS/BPM using $\hat{\sigma}^2$ from (13) and estimating $\{\gamma_k^2\}_{k=1}^n$ as in Section 4.1 with the true $\mathbf{\Psi}$; (e) Empirical BOSS/BPM using $\hat{\sigma}^2$ from (13) and estimating $\{\gamma_k^2\}_{k=1}^n$ as in Section 4.1 with $\mathbf{\Psi}$ consisting of the first $r = 3$ discrete Fourier series from (16); (f) Empirical BOSS/BPM using $\hat{\sigma}^2$ from (13) and estimating $\{\gamma_k^2\}_{k=1}^n$ as in Section 4.2 with λ chosen using GCV [11].

For all empirical Bayes methods we use $b_f = 0.1$ in (14) to compute the lower threshold b for $\{\hat{\gamma}_k^2\}_{k=1}^n$, such that all $\hat{\gamma}_k^2 < b$ from (17), (18) are set to b instead. In simulations not detailed here we observed that the exact choice of b_f is not critical for the performance.

In Figure 1 we study the model order selection performances and in Figure 2 we study the coefficient estimation performances. We observe that the empirical BOSS/BPM based methods give a higher performance than AICc and BIC does (generally about 0.5 to 1 dB better, and sometimes significantly more). Also, the differences between the different empirical Bayesian methods are very small. By estimating σ^2, $\{\gamma_k^2\}_{k=1}^n$ we lose around 0.5 to 1 dB from the performance of the optimal BOSS/BPM where these parameters are known. We have obtained similar results from many other numerical examples which are omitted here due to space constraints.

6. CONCLUSIONS

We have studied linear regression with an unknown model order assuming zero-mean Gaussian noise and a zero-mean Gaussian distributed coefficient vector. Under this model we have derived empirical Bayesian versions of the maximum *a posteriori* probability model order selector and the MMSE coefficient vector estimate. These empirical Bayesian methods have been shown to outperform the classical approaches AICc and BIC, both in model selection and in coefficient estimation examples. Since our methods have low computational complexity and show very good performance, we consider them attractive alternatives in the context of the model (2).

Code (in Matlab) for the methods presented here is available at `www.it.uu.se/katalog/ys/software/empBOSS_BPM`.

7. REFERENCES

[1] F. Gustafsson and H. Hjalmarsson, "Twenty-one ML Estimators for Model Selection," *Automatica*, vol. 31, pp. 1377–1392, 1995.

[2] H. Hjalmarsson and F. Gustafsson, "Composite modeling of transfer functions," *IEEE Transactions on Automatic Control*, vol. 40, pp. 820–832, May 1995.

[3] E. G. Larsson and Y. Selén, "Linear regression with a sparse parameter vector," *IEEE Transactions on Signal Processing*, 2006. To Appear: http://www.ee.kth.se/php/modules/publications/reports/2006/IR-EE-KT_2006_013.pdf.

[4] Y. Selén and E. G. Larsson, "Parameter estimation and order selection for linear regression problems," in *14th EUSIPCO*, (Florence, Italy), September 4 – 8 2006.

[5] L. Wasserman, "Bayesian model selection and model averaging," *Journal of Mathematical Psychology*, vol. 44, pp. 92–107, March 2000.

[6] P. Stoica and Y. Selén, "Model-order selection — A review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, pp. 36–47, July 2004.

[7] T. Söderström and P. Stoica, *System Identification*. London, UK: Prentice Hall International, 1989.

[8] C. Févotte and S. J. Godsill, "Sparse linear regression in unions of bases via Bayesian variable selection," *IEEE Signal Processing Letters*, vol. 13, pp. 441–444, July 2006.

[9] A. A. M. Saleh and R. A. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE Journal on Selected Areas in Communications*, vol. 5, pp. 128–137, February 1987.

[10] A. E. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet. Math. Dokl.*, vol. 4, pp. 1035–1038, 1963.

[11] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM Review*, vol. 34, pp. 561–580, December 1992.

[12] J. E. Cavanaugh, "Unifying the derivations for the Akaike and corrected Akaike information criteria," *Statistics and Probability Letters*, vol. 33, pp. 201–208, April 1977.

[13] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.