14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

PARAMETER ESTIMATION AND ORDER SELECTION FOR LINEAR REGRESSION PROBLEMS

Yngve Selén and Erik G. Larsson

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

KTH/EE Communication Theory Laboratory
Royal Institute of Technology, Osquldas väg 10
SE-100 44 Stockholm, Sweden.
email: erik.larsson@ee.kth.se

ABSTRACT

Parameter estimation and model order selection for linear regression models are two classical problems. In this article we derive the minimum mean-square error (MMSE) parameter estimate for a linear regression model with unknown order. We call the so-obtained estimator the Bayesian Parameter estimation Method (BPM). We also derive the model order selection rule which maximizes the probability of selecting the correct model. The rule is denoted BOSS—Bayesian Order Selection Strategy. The estimators have several advantages: They satisfy certain optimality criteria, they are non-asymptotic and they have low computational complexity. We also derive "empirical Bayesian" versions of BPM and BOSS, which do not require any prior knowledge nor do they need the choice of any "user parameters". We show that our estimators outperform several classical methods, including the AIC and BIC for order selection.

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 parameter 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 $n_{\min} \leq k \leq n$, $\boldsymbol{X}_k = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_k]$ (i.e., $\boldsymbol{X}_k$ consists of the first k columns of $\boldsymbol{X}$), and $\boldsymbol{h}_k = [h_1 \cdots h_k]^T$. The model order k is assumed to be unknown. We consider two interrelated tasks: the problem of estimating $\boldsymbol{h}$ given $\boldsymbol{X}$ and $\boldsymbol{y}$, and the problem of detecting the order k, given $\boldsymbol{X}$ and $\boldsymbol{y}$.

If $n \leq N$, $\boldsymbol{X}$ were full rank, and k were known, then $\boldsymbol{h}_k$ could be estimated by the maximum-likelihood (ML) method. It is well known [1] that the ML estimator coincides with the least squares (LS) estimator if the noise $\boldsymbol{\epsilon}$ is white and Gaussian:

$$\hat{\boldsymbol{h}}_{k,\mathrm{ML}} = \hat{\boldsymbol{h}}_{k,\mathrm{LS}} = (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^T \boldsymbol{y}. \tag{3}$$

However, this ML estimate is not directly useful if k is unknown. For example, by using a value of k in (3) which is larger than the true model order, then one would estimate many parameters which were in fact equal to zero. This would result in a suboptimal estimate with a large variance. On the contrary, if one used a value of k which were smaller than the true order, then the estimate obtained by (3) would be biased.

In this article we obtain the best possible bias-variance tradeoff for the estimation of $\boldsymbol{h}$, in the sense of minimum mean squared error (MMSE). We also solve the model order selection problem by obtaining the model order k with the highest posterior probability; this selection rule is optimal in the sense that no other method can have a higher probability of picking the correct order. To facilitate our derivation, we will assume that the parameters h_j are independent, Gaussian random variables: $h_j \sim \mathcal{N}(0, \gamma_j^2)$. In other words, $\boldsymbol{h}_k \sim \mathcal{N}(\boldsymbol{0}, \mathrm{diag}[\boldsymbol{\gamma}_k])$ where $\boldsymbol{\gamma}_k = [\gamma_1^2, \cdots, \gamma_k^2]^T$. The Gaussian assumption is convenient from an analytical point of view. Additionally, Gaussian random variables have the largest uncertainty (in the sense of entropy) for a given variance [2]. This makes the Gaussian density a natural choice for the prior distributions. To begin, we will assume that the variances $\{\gamma_j^2\}_{j=1}^n$ and the noise variance σ^2 are known. Then, in Section 4, we will show how to use our method without knowing any of the variances.

Note: The models $\{\mathcal{M}_k\}_{k=n_{\min}}^n$ considered in this article are common in signal processing applications, including finite-impulse-response (FIR) filter identification and the estimation of polynomial coefficients [1]. By way of contrast, in statistical data analysis problems it is common to consider *sparse* models, where the true model can contain *any* subset of the regressors $\{\boldsymbol{x}_j\}_{j=1}^n$ [3]. The sparse regression problem is related to the problem we study in this article, and in some sense also more difficult since the number of possible models grows exponentially with n, rather than linearly (which is the case for the model we study in this paper). In a related paper [4], we derived solutions (albeit approximative) to the parameter estimation and model selection problems for a sparse regression model. This article was inspired by our developments in [4]. The main contrasts to [4] are that (i) for the parameter estimation and order selection problems considered in this paper, no approximations are necessary; (ii) owing to the special structure of the problem, we can obtain computationally very efficient estimates; and (iii) we use a more flexible model

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

for the parameter vector (in particular, we can let the variances of distinct coefficients be different).

1.2 Background and Related Work

The problems of model order selection and parameter estimation with unknown model order are as old as the idea of parametric modeling. Traditionally, the parameter estimation problem has been split in two steps: First, the model order selection problem is solved, and then the parameter vector is estimated assuming that the selected order is equal to the true one. This is generally not optimal. In, e.g., [5], [6], the concept of Bayesian model averaging is discussed. In this type of approach, the quantity of interest (e.g., a prediction or a parameter value) is first computed separately for each of the considered models. A weighted sum, with the model posterior probabilities as weights, is then used for the final model averaged computation of the quantity of interest. The estimator we develop in Section 2 (see, e.g., (5)) is a type of Bayesian model average. In [3], Bayesian model averaging was considered for a different set of linear regression models under some other *a priori* assumptions.

Model order selection is the art of determining the "best" model order from a set of model order candidates. Usually the goal is to maximize the probability that the selected model is the "true" one. Model order selection procedures for which the probability of selecting the correct order approaches one (assuming such a correct model order exists among the candidates) as the number of data samples increases are called *consistent.*

A commonly used group of model selection criteria is the *information criteria* (IC) group. There exist many different IC [8], but the most frequently used are the information criterion of Akaike (AIC) [9] (which is not consistent) and the Bayesian Information Criterion (BIC) [10] (which is consistent). The IC have a common form: they select the model which minimizes

$$-2\ln p(\boldsymbol{y}|\hat{\boldsymbol{h}}_{k,\mathrm{ML}}) + k\cdot\eta(\cdot) \tag{4}$$

where $p(\boldsymbol{y}|\hat{\boldsymbol{h}}_{k,\mathrm{ML}})$ is the probability density function (pdf) for the measured data given the ML parameter vector estimate of order k. The first term in (4) is a measure of how well the model fits the data; by adding more parameters the model will be a better fit and the value of this term will decrease. For the linear regression problem (2), the first term equals [11]:

$$-2\ln p(\boldsymbol{y}|\hat{\boldsymbol{h}}_{k,\mathrm{ML}}) = N\ln\hat{\sigma}_k^2 + \text{constant}$$

where $\hat{\sigma}_k^2 = \|\boldsymbol{y} - \boldsymbol{X}_k\hat{\boldsymbol{h}}_{k,\mathrm{ML}}\|^2/N$ is the ML estimate of the noise variance. The second term in (4), $k\cdot\eta(\cdot)$, is usually called a "penalty" term and it increases with increasing k. (The function $\eta(\cdot)$ is specific to the particular IC being used; it typically depends on k and N.) The role of this penalty is to penalize models with many parameters (which have a low value of $-2\ln p(\boldsymbol{y}|\hat{\boldsymbol{h}}_{k,\mathrm{ML}})$). In other words, it ensures that a simple model is preferred unless a more complex model fits the data significantly better. The function $\eta(\cdot)$ takes on the values $\eta = 2$ for AIC and $\eta(N) = \ln N$ for BIC. All information criteria we are aware of, and which fit into this framework, are asymptotic in the sense that they are derived assuming $N \to \infty$. However, there exists a bias-corrected small-sample version of AIC—AIC$_\mathrm{c}$ [12]—which is applicable for certain types of models. It uses $\eta(k, N) = 2N/(N - k - 2)$.

More detailed reviews of IC-based model selection criteria can be found in [11], [13].

1.3 Contribution of This Work

In this article, we derive the MMSE parameter estimate $\hat{\boldsymbol{h}}_{\mathrm{MMSE}}$ for the linear regression problem with unknown model order. We denote the resulting estimator the BPM (Bayesian Parameter estimation Method). We also derive the model order estimator which has the largest probability of returning the correct answer. This model order estimator will be called BOSS (Bayesian Order Selection Strategy). The estimators (see (8) and (12)) have a number of advantages:

1. **Optimality:** Our BPM parameter estimate is optimal in the MMSE sense. BOSS is also optimal: no other order selection algorithm will pick the correct order more often on the average. This optimality is valid if all assumptions used in the derivations hold.
2. **Short-sample performance:** Many model order selection algorithms (notably, the IC) are asymptotic. BPM and BOSS are not, and therefore they show good performance also for short data sequences.
3. **Computational efficiency:** BPM and BOSS can be efficiently computed.
4. **No user parameters:** We present versions of our methods which can be used without having any *a priori* knowledge of the data or selecting any user parameters. (See Section 4.)

We remark upon the fact that neither BOSS nor BPM are completely new estimators. Although not explicitly stated, they can be easily obtained, e.g., from the results in [7]. In this light, the new contributions in this article over [7] are (1) an, in our opinion, "cleaner" derivation; (2) the empirical approach described in Section 4 for treating unknown *a priori* parameters; (3) the numerical study in Section 5.

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

In this section we describe the BPM (Bayesian Parameter estimation Method) by deriving the MMSE parameter estimate of $\boldsymbol{h}$ under the assumption that one of the models $\{\mathcal{M}_k\}_{k=n_{\min}}^{n}$ in (2) generated the data. We also assume that the parameter vector elements are zero mean Gaussian with known variances $\{\gamma_j^2\}_{j=1}^{n}$, and that the noise variance σ^2 is known. (These assumptions will be relaxed in Section 4.) The MMSE estimate is equal to the conditional mean,

$$\hat{\boldsymbol{h}}_{\mathrm{MMSE}} = E[\boldsymbol{h}|\boldsymbol{y}] = \sum_{k=n_{\min}}^{n} P(\mathcal{M}_k|\boldsymbol{y})E[\boldsymbol{h}|\boldsymbol{y},\mathcal{M}_k]. \tag{5}$$

Using Bayes' rule we obtain

$$\hat{\boldsymbol{h}}_{\mathrm{MMSE}} = \sum_{k=n_{\min}}^{n} P(\mathcal{M}_k)\frac{p(\boldsymbol{y}|\mathcal{M}_k)}{p(\boldsymbol{y})}E[\boldsymbol{h}|\boldsymbol{y},\mathcal{M}_k] \tag{6}$$

where

$$p(\boldsymbol{y}) = \sum_{k=n_{\min}}^{n} p(\boldsymbol{y}, \mathcal{M}_k) = \sum_{k=n_{\min}}^{n} P(\mathcal{M}_k) p(\boldsymbol{y}|\mathcal{M}_k). \quad (7)$$

Combining (6) and (7) we obtain the Bayesian Parameter estimation Method,

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

Generally (as is commonly done [6], [8]), if nothing else 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:

$$P(\mathcal{M}_k) = \frac{1}{n - n_{\min} + 1}, \qquad k = n_{\min}, \ldots, n. \quad (9)$$

What remains to evaluate (8) is then to compute $p(\boldsymbol{y}|\mathcal{M}_k)$ and $E[\boldsymbol{h}|\boldsymbol{y}, \mathcal{M}_k]$.

Computation of $p(\boldsymbol{y}|\mathcal{M}_k)$: 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

$$\boldsymbol{Q}_k = \boldsymbol{X}_k \boldsymbol{\Gamma}_k \boldsymbol{X}_k^T + \sigma^2 \boldsymbol{I}$$

where we have used $\boldsymbol{\Gamma}_k = \text{diag}[\boldsymbol{\gamma}_k]$. So,

$$p(\boldsymbol{y}|\mathcal{M}_k) = \frac{1}{\sqrt{2\pi}^N} \frac{1}{|\boldsymbol{Q}_k|^{1/2}} \exp\left(-\frac{1}{2} \boldsymbol{y}^T \boldsymbol{Q}_k^{-1} \boldsymbol{y}\right). \quad (10)$$

Computation of $E[\boldsymbol{h}|\boldsymbol{y}, \mathcal{M}_k]$: Clearly, assuming 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 (Theorem 10.2 of [14], for example), the conditional mean evaluates to

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

We now have all the ingredients necessary to compute (8): namely (9), (10) and (11).

3. OPTIMAL MODEL ORDER SELECTION

The selection of the "best" model order is a classical problem. Here we derive BOSS (Bayesian Order Selection Strategy)—our method for model order selection. Applying Bayes' Theorem, as in the previous section, 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

$$\begin{aligned} \text{BOSS}: \quad \hat{k} &= \arg \max_{k=n_{\min},\ldots,n} P(\mathcal{M}_k|\boldsymbol{y}) \\ &= \arg \max_{k=n_{\min},\ldots,n} P(\mathcal{M}_k) p(\boldsymbol{y}|\mathcal{M}_k) (12) \end{aligned}$$

which can be computed using (9) and (10) from the previous section. One can show (see, e.g., the discussion around Equation (37) in [11]) that the order with the highest *a posteriori* probability is also the choice which is the most likely to equal the correct order.

4. EMPIRICAL BAYESIAN ESTIMATORS

BPM (8) and BOSS (12) require knowledge of the *a priori* parameters $\{P(\mathcal{M}_k)\}_{k=n_{\min}}^n$, $\{\gamma_j^2\}_{j=1}^n$ and σ^2. In practice, these parameters may not be known perfectly. In this section, we present an efficient, pragmatic solution to the problem of using BPM and BOSS when the *a priori* parameters are completely unknown. For simplicity, we restrict the discussion to the case that the h_j are i.i.d., so that $\gamma_j^2 = \gamma^2$, $\forall\ j$. Furthermore, as in (9), we make the common assumption [6], [8] that, if nothing at all is known about the probabilities of the considered orders *a priori*, all model orders are equally probable. The relevant parameters are then γ^2, σ^2.

Our approach is to estimate the parameters γ^2, σ^2 from the data. This will result in a so-called "empirical Bayesian" method [15]. Note that, in a strict sense, estimation of γ^2, σ^2 from the data voids the optimality of our estimators (8), (12). This is so because the parameters γ^2, σ^2 are *a priori* parameters, and should therefore not depend on the observation vector $\boldsymbol{y}$. Nevertheless, estimation of γ^2, σ^2 from the data appears to be an attractive, pragmatic way of handling the situation when these parameters are completely unknown. (Note: A different approach could be to assume that $\{P(\mathcal{M}_k)\}_{k=n_{\min}}^n$, $\{\gamma_j^2\}_{j=1}^n$ and σ^2 are random with known distributions; see, e.g., [3].)

An unbiased, consistent estimate of σ^2 can be obtained by taking [1]

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

where $\hat{\boldsymbol{h}}_{n,\text{ML}}$ is obtained from (3).

Next, under the model $\mathcal{M}_k$, γ^2 can be estimated by

$$\hat{\gamma}^2_{\mathcal{M}_k} = \frac{\|\hat{\boldsymbol{h}}_{k,\text{ML}}\|^2}{k}.$$

By weighting the above estimates with $P(\mathcal{M}_k)$, we can devise the following estimator[1]:

$$\hat{\gamma}^2 = \sum_{k=n_{\min}}^{n} \frac{\|\hat{\boldsymbol{h}}_{k,\text{ML}}\|^2}{k} P(\mathcal{M}_k). \quad (14)$$

We obtain the empirical Bayesian versions of our estimators by inserting (13) and (14) into the expressions (8) and (12).

[1] Note that estimation of γ^2 under $\mathcal{M}_0$ is an ill-posed problem. The case $n_{\min} = 0$ can thus not be handeled by our empirical Bayesian estimators.

5. NUMERICAL EXAMPLES

We evaluate the performance of our methods by means of Monte-Carlo simulations. For the evaluation of our parameter estimator BPM (8) we use the empirical MSE of the parameter 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 parameter values for realization number m, and M is the total number of Monte-Carlo runs. For the evaluation of our model order estimator BOSS (12) we use the percentage of correctly estimated orders: $M^{-1}\sum_{m=1}^{M}\delta(\hat{k}^{(m)}-k^{(m)})\cdot 100$ [%], where $\delta(t)$ denotes the discrete-time unit impulse and $\hat{k}^{(m)}$ and $k^{(m)}$ denote the estimated and true orders for realization number m, respectively. We choose $M = 10000$.

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 $n = 10$ (these are also the values of $n_{\min}, n$ supplied to the estimators). We try both long ($N = 300$) and short ($N = 30$) data sequences. Our regressor matrix $\boldsymbol{X}$ is composed of i.i.d. $\mathcal{N}(0,1)$ elements.

We compare the performance of our estimators (both the true Bayesian—with all prior knowledge available—and the empirical Bayesian) with the performance obtained using BIC, and AIC (for $N = 300$) or AIC$_\mathrm{c}$ (for $N = 30$). In the parameter estimation examples with AIC/AIC$_\mathrm{c}$ and BIC, these are first used to obtain an order estimate $\hat{k}$ (by minimizing (4)); $\boldsymbol{h}$ is then estimated using ML (3) for the order $\hat{k}$.

5.1 Example 1

We first evaluate the performance of our methods for noise variances from 10 dB down to −20 dB. The variances of the true parameter values $\{h_j\}$ are set to $\gamma^2 = 1$.

Figures 1, 2 show the results. Here BPM (8) and its empirical version consistently outperform the other methods. Also, the performance difference between BPM and empirical BPM is small. For $N = 300$ all estimators but AIC have very similar performance, except when the noise variance is large.

In Figure 3 we show the order selection results. Our estimator, BOSS (12), consistently outperforms all other methods. The empirical version of BOSS performs very close to the true Bayesian version.

5.2 Example 2 (parameter mismatch)

Next, we investigate the estimators' sensitivity to a mismatch in the *a priori* parameters γ^2, σ^2. For brevity, we only show the results for order selection and for $N = 30$. Similar conclusions hold for the parameter estimation and the order selection results for $N = 300$, but the differences between the methods are less pronounced.

In Figure 4, we supply our Bayesian estimator (12) with a mismatched $\gamma^2 = 0$ dB, while the actual γ^2 used to generate the data is varied from −10 dB to 10 dB. The true $\sigma^2 = -10$ dB is given to BOSS. (The empirical BOSS estimates γ^2, σ^2.) As can be seen from the figure, BOSS is not very sensitive to mismatched γ^2-values.

In Figure 5 we supply BOSS with $\sigma^2 = -10$ dB, when the true σ^2 is varied from 10 dB down to −20 dB. The true $\gamma^2 = 1$ is given to the estimator (the empirical BOSS estimates γ^2, σ^2). As can be seen, as long as the σ^2 provided to BOSS does not deviate too much from the true value, the estimator retains very good performance. Fortunately, σ^2 is relatively easy to estimate accurately from the data using (13).

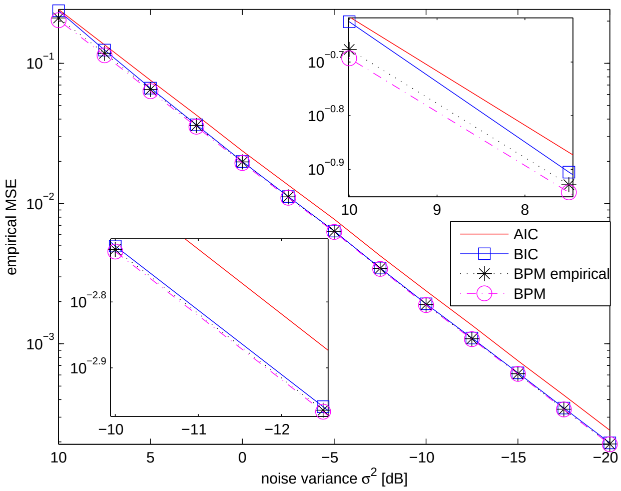

Figure 1: Parameter estimation: Long data sequence, $N = 300$. (The small plots are closeups.)

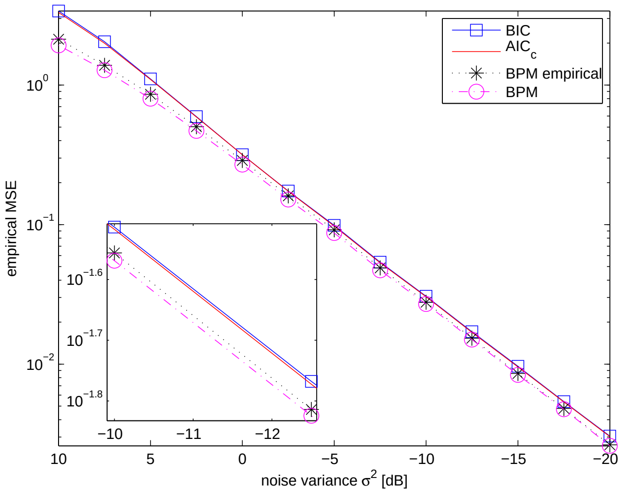

Figure 2: Parameter estimation: Short data sequence, $N = 30$. (The small plot is a closeup.)

This example indicates that neither BPM nor BOSS is particularly sensitive to the choice of γ^2, σ^2. This may be a contributing reason for why the empirical Bayesian variants of our estimators work so well.

6. CONCLUSIONS

We have derived the optimal parameter estimator, in the MMSE sense, for linear regression when model order is unknown. We have also derived the optimal model order estimator for the regression problem. Our estimators[2], denoted BPM and BOSS, respectively, possess a number of advantages over classical approaches (see Section 1.3). Since they are Bayesian, they require knowledge of some *a priori* parameters. However, we have also derived empirical Bayesian variants of BPM and BOSS which do

[2] Free implementations of our estimators (in Matlab) can be obtained from [16].

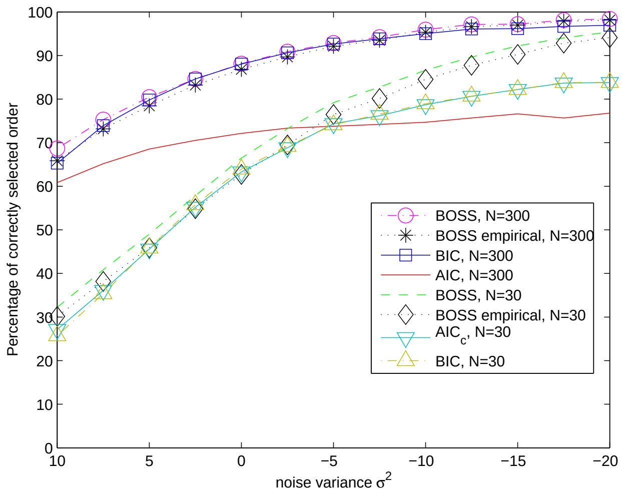

Figure 3: Model order selection.

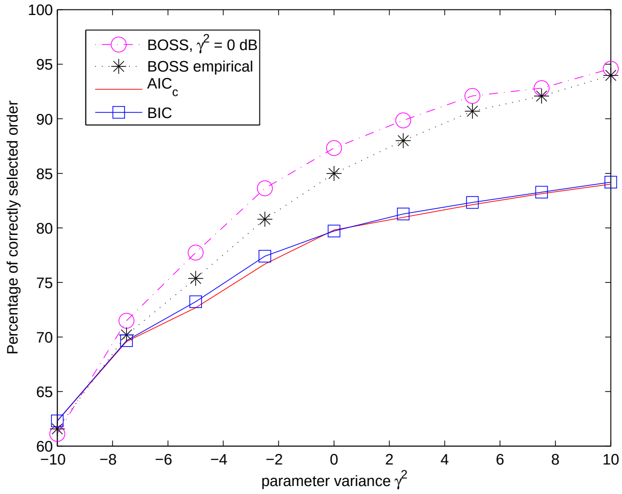

Figure 4: Order selection: Mismatching γ^2, $N = 30$.

not require the choice of any user parameters.

In numerical examples, both for short and long data sequences, we demonstrate that our estimators outperform the classical approaches AIC, AIC$_c$ and BIC. BPM and BOSS are relatively insensitive to the choice of the *a priori* parameters. Also, their empirical Bayesian variants work very well. Since our estimators have low complexity, and since they possess certain optimality properties, we believe that they should be considered attractive alternatives for estimation and detection in the context of the model (2).

REFERENCES

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

[2] E. T. Jaynes, *Probability theory: the logic of science*. Cambridge, UK: Cambridge University Press, 2003.

[3] A. E. Raftery, D. Madigan, and J. A. Hoeting, "Bayesian model averaging for linear regression models," *Journal of the American Statistical Association*, vol. 92, pp. 179–191, March 1997.

[4] E. G. Larsson and Y. Selén, "Linear Regression With a Sparse Parameter Vector," in *IEEE ICASSP*, (Toulouse, France), May 14 – 19 2006. To Appear: [www.s3.kth.se/~elarsso/preprints].

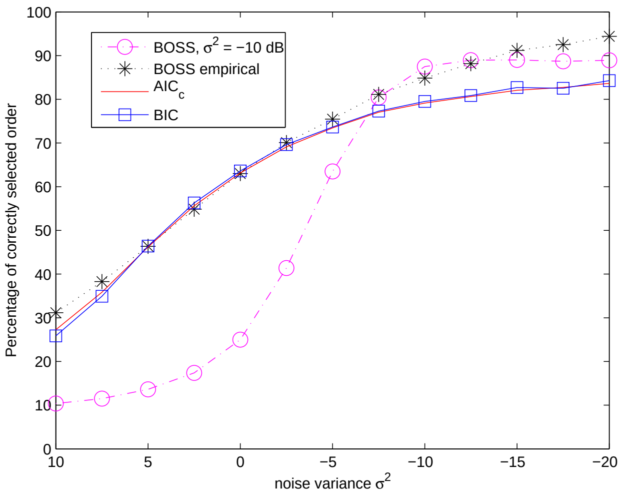

Figure 5: Order selection: Mismatching σ^2, $N = 30$.

[5] J. A. Hoeting, D. Madigan, A. E. Raferty, and C. T. Volinsky, "Bayesian Model Averaging: A Tutorial," *Statistical Science*, vol. 14, pp. 382–417, November 1999.

[6] L. Wasserman, "Bayesian Model Selection and Model Averaging," *Journal of Mathematical Psychology*, vol. 44, pp. 92–107, March 2000.

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

[8] K. P. Burnham and D. R. Anderson, *Model Selection and Multi-Model Inference*. New York: Springer, 2002.

[9] H. Akaike, "A New Look at Statistical Model Identification," *IEEE Trans. Automatic Control*, vol. 19, pp. 716–723, 1974.

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

[11] 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.

[12] C. Hurvich and C. Tsai, "A Corrected Akaike Information Criterion for Vector Autoregressive Model Selection," *J. Time Series Anal.*, vol. 14, pp. 271–279, 1993.

[13] A. D. Lanterman, "Schwarz and Wallace and Rissanen: Intertwining themes in theories of model order estimation," *International Statistical Review*, vol. 69, pp. 185–212, August 2001.

[14] S. M. Kay, *Fundamentals of Statistical Signal Processing, Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[15] E. I. George and D. P. Foster, "Calibration and empirical Bayes variable selection," *Biometrika*, vol. 87, pp. 731–747, December 2000.

[16] `http://user.it.uu.se/~ys/software/BPM_BOSS/`