2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Kernel Selection in Linear System Identification

Part II: A Classical Perspective

Tianshi Chen, Henrik Ohlsson, Graham C. Goodwin and Lennart Ljung

***Abstract*— In this companion paper, the choice of kernels for estimating the impulse response of linear stable systems is considered from a classical, "frequentist", point of view. The kernel determines the regularization matrix in a regularized least squares estimate of an FIR model. The quality is assessed from a mean square error (MSE) perspective, and measures and algorithms for optimizing the MSE are discussed. The ideas are tested on the same data bank as used in Part I of the companion papers. The resulting findings and conclusions in the two papers are very similar despite the different perspectives.**

I. Introduction

We study the problem of estimating the impulse response of linear stable systems. Consider a single-input–single-output linear stable system

$$y(t) = G_0(q)u(t) + v(t) \tag{1}$$

Here, $y(t)$ is the measured output, q is the shift operator, $qu(t) = u(t+1)$, $v(t)$ is the additive white noise, independent of the input $u(t)$, and the transfer function is

$$G_0(q) = \sum_{k=1}^{\infty} g_k^0 q^{-k} \tag{2}$$

The coefficients $\{g_k^0\}_{k=1}^{\infty}$ form the *impulse response* of the system. Given the input-output data $\{u(t), y(t), t = 1, \ldots, N\}$, the goal is to find estimates $\{\hat{g}_k\}_{k=1}^{\infty}$ of the impulse response coefficients $\{g_k^0\}_{k=1}^{\infty}$.

This is of course a problem that has been studied for a long time, and has a huge literature, see *e.g.,* [1]. Recently, this problem is studied from a Gaussian process regression perspective [2], [3], [4]. In particular, in Part I of the companion papers [4], it is discussed how to find suitable kernels for the Gaussian process regression that give estimates $\{\hat{g}_k\}_{k=1}^{\infty}$ as good as possible. On the other hand, in our recent paper [5], we formulate a classical regularization approach, focused on finite impulse response (FIR) models, and show that this basic regularized least squares approach is a focal point for interpreting other approaches, like Bayesian inference and Gaussian process regression. In this contribution, we continue our discussions and focus on how to deal with the kernel (regularization matrix) selection problem from a classical, "frequentist", perspective.

In [4], the quality of the kernel is assessed from its associated marginal likelihood. In contrast, the quality of the regularization matrix (kernel) is assessed here from its associated mean square error (MSE). The ideas are tested on the same data bank as used in Part I of the companion papers [4]. The resulting findings and conclusions in the two papers are very similar despite the different perspectives.

Tianshi Chen, Henrik Ohlsson and Lennart Ljung are with the Department of Electrical Engineering, Linköping University, Linköping, Sweden {tschen,ohlsson,ljung}@isy.liu.se

Graham C. Goodwin is with the School of Electrical Engineering and Computer Science, The University of Newcastle, Newcastle, Australia Graham.Goodwin@newcastle.edu.au

II. Preliminary

Before running into the details, we first provide some preliminary discussions that will be used later.

A. Model Structures

Traditionally, the problem of estimating impulse responses of linear systems is approached by selecting a particular parametrization (or model structure) of the impulse response:

$$y(t) = G(q, \theta)u(t) + v(t) \tag{3}$$

where

$$G(q, \theta) = \sum_{k=1}^{\infty} g_k(\theta) q^{-k} \tag{4}$$

The finite-dimensional parameter θ is estimated, *e.g.,* as

$$\hat{\theta}_N = \arg\min_{\theta} \sum_{t=1}^{N} (y(t) - G(q, \theta)u(t))^2 \tag{5}$$

and then the impulse response estimate is found as

$$\hat{g}_k = g_k(\hat{\theta}_N), \qquad k = 1, 2, \ldots, \infty \tag{6}$$

This is a simple description of the prediction error method (PEM) which gives the maximum likelihood (ML) estimate when the noise $v(t)$ is Gaussian, see, *e.g.,* [1].

B. Classical Estimation Goal

In the classical perspective, the goal is to find estimates $\{\hat{g}_k\}_{k=1}^{\infty}$ of $\{g_k^0\}_{k=1}^{\infty}$ such that

$$\sum_{k=1}^{\infty} (g_k^0 - \hat{g}_k)^2 \tag{7}$$

is as small as possible.

Now, the estimates $\{\hat{g}_k\}_{k=1}^{\infty}$ will be random variables, since they are formed from data $\{u(t), y(t), t = 1, \ldots, N\}$ that is affected by the noise $v(t)$, so the above sum (7) is a random variable. Therefore we take expectation w.r.t. the noise $v(t)$ to form the *mean square error*

$$MSE(\hat{\theta}_N) = \sum_{k=1}^{\infty} E(g_k^0 - \hat{g}_k)^2 \tag{8}$$

978-1-61284-799-3/11/$26.00 ©2011 IEEE

III. Regularized Least Squares

An especially simple choice of model structure (4) is to truncate the impulse response, and let the parameter vector be the impulse response coefficients themselves. This is known as an FIR (finite impulse response) model:

$$G(q,\theta)=\sum_{k=1}^{n} g_k q^{-k}, \quad \theta=\begin{bmatrix} g_1 & g_2 & \dots & g_n \end{bmatrix}^T \tag{9}$$

We will from now on assume that $g_k^0=0, k>n$ so that the true unknown system can be described as an FIR model. In the following, we let the true impulse response coefficients be denoted by

$$\theta_0=\begin{bmatrix} g_1^0 & g_2^0 & \dots & g_n^0 \end{bmatrix}^T \tag{10}$$

The problem of estimating the FIR model (9) can be written as a linear regression as follows:

$$\begin{aligned} Y_N &= \begin{bmatrix} y(n+1) & \dots & y(N) \end{bmatrix}^T \\ \Phi_N &= \begin{bmatrix} u(n) & u(n+1) & \dots & u(N-1) \\ u(n-1) & u(n) & \dots & u(N-2) \\ \vdots & \vdots & \dots & \vdots \\ u(1) & u(2) & \dots & u(N-n) \end{bmatrix} \\ V_N &= \begin{bmatrix} v(n+1) & \dots & v(N) \end{bmatrix}^T \\ Y_N &= \Phi_N^T\theta + V_N \end{aligned} \tag{11}$$

Note that eq. (11) corresponds to eq. (3) in [4].

A. Least Squares

Corresponding to (10), let the well-known least squares (LS) estimate of θ_0 be denoted by

$$\hat{\theta}_N^{LS}=\begin{bmatrix} \hat{g}_1^{LS} & \hat{g}_2^{LS} & \dots & \hat{g}_n^{LS} \end{bmatrix}^T \tag{12a}$$

which is given by

$$\hat{\theta}_N^{LS}=\arg\min_{\theta}\|Y_N-\Phi_N^T\theta\|^2=R_N^{-1}\Phi_N Y_N \tag{12b}$$

$$R_N=\Phi_N\Phi_N^T \tag{12c}$$

For FIR models of high order n (say 125) this estimate will typically have large variance.

B. Regularized Least Squares: Bias–Variance Trade-Off

The classical way of handling high variance estimates, is to allow some bias in the estimate that reduces the variance, but reaches a smaller MSE (MSE is the sum of the square of the bias and the variance). For linear regressions, the standard way is to introduce *regularization*. Corresponding to (10), let the regularized LS estimate of θ_0 be denoted by

$$\hat{\theta}_N^R=\begin{bmatrix} \hat{g}_1^R & \hat{g}_2^R & \dots & \hat{g}_n^R \end{bmatrix}^T \tag{13}$$

which is given by

$$\begin{aligned} \hat{\theta}_N^R &= \min_{\theta}\|Y_N-\Phi_N^T\theta\|^2+\theta^T Z^{-1}\theta, \quad Z^{-1}\geq 0 \\ &= (R_N+Z^{-1})^{-1}\Phi_N Y_N \end{aligned} \tag{14}$$

where Z^{-1} is the *regularization matrix*. The regularized estimate $\hat{\theta}_N^R$ depends on Z but we suppress this in the notation. Note that this estimate $\hat{\theta}_N^R$ corresponds to eq. (19) in [4] and the matrix Z^{-1} corresponds to the kernel $\sigma^2(\hat{\lambda}^2\hat{K}(\hat{\beta}))^{-1}$ in eqs. (16) and (19) in [4].

C. Mean Square Error

From (11) and (10), the true system (3) can be written as

$$Y_N=\Phi_N^T\theta_0+V_N \tag{15}$$

Then the mean square error matrix of $\hat{\theta}_N^R$ is

$$\begin{aligned} M_N(\hat{\theta}_N^R) &= E(\hat{\theta}_N^R-\theta_0)(\hat{\theta}_N^R-\theta_0)^T \\ &= (R_N+Z^{-1})^{-1}(\sigma^2 R_N+Z^{-1}\theta_0\theta_0^T Z^{-T})(R_N+Z^{-1})^{-1} \end{aligned} \tag{16}$$

Consequently, the MSE measure (8) for the regularized estimate $\hat{\theta}_N^R$ is

$$MSE(\hat{\theta}_N^R)=\operatorname{trace} M_N(\hat{\theta}_N^R) \tag{17}$$

We also see how Z affects the bias variance trade-off. Roughly speaking, the larger Z (the smaller Z^{-1}), the smaller the bias will be but the larger the variance. In the limiting case $Z^{-1}=0$ we are back in the un-regularized case (12b). The matrix $M_N(\hat{\theta}_N^R)$ will be our main tool to evaluate the quality aspects of various choices for Z.

Remark 3.1: To compute the mean square error of the estimate we make the follow assumptions:

1) The disturbance $v(t)$ is white noise with variance σ^2.
2) The input u is a known sequence; hence R_N is a known, deterministic matrix.
3) The regularization matrix Z is a known, constant matrix; hence independent of V_N.

Actually, we will work with cases later on ("empirical Bayes") where Z is partly estimated from data. Then assumption 3 does not hold strictly. But then Z will converge to an V-independent matrix (as $N\to\infty$), so the expressions will hold asymptotically. Otherwise the expressions above are exact, and not asymptotic in N.

IV. Regularization Matrix (Kernel) Selection

A. A Matrix Inequality for the MSE

Let $Q_0=\theta_0\theta_0^T$. Then $M_N(\hat{\theta}_N^R)$ in (16) can be rewritten as

$$\begin{aligned} M_N(\hat{\theta}_N^R) &= M_N(R_N,Q_0,Z) \\ &= (R_N+Z^{-1})^{-1}(\sigma^2 R_N+Z^{-1}Q_0 Z^{-1})(R_N+Z^{-1})^{-1} \end{aligned} \tag{18}$$

The following algebraic matrix relationship is important:

$$M_N(R_N,Q_0,Z)\geq M_N(R_N,Q_0,Q_0/\sigma^2)\ \forall\ R_N>0, Q_0, Z\geq 0 \tag{19}$$

Note that the inequality holds in a matrix sense, *i.e.,* $M_N(R_N,Q_0,Z)-M_N(R_N,Q_0,Q_0/\sigma^2)$ is positive semi-definite for all positive definite R_N and positive semi-definite Q_0, Z. So the mean square error matrix $M_N(R_N,Q_0,Z)$ is minimized by the choice $Z=Q_0/\sigma^2$. The proof consists of elementary matrix calculations and can be found in [5].

Remark 4.1: It may happen that Z may be singular, so the expressions in (16) contain non-existing inverses. If so, expressions like $(R_N+Z^{-1})^{-1}$ are rewritten as

$$(R_N+Z^{-1})^{-1} = (ZR_N+I)^{-1}Z \tag{20}$$

to contain well-defined expressions.

Remark 4.2: If Z in (14) has rank 1, the regularization is best interpreted through the solution:

$$\hat{\theta}_N^R = (R_N+Z^{-1})^{-1}\Phi_N Y_N = (ZR_N+I)^{-1}Z\Phi_N Y_N \tag{21}$$

Assume Z=LL^T for a column vector L. Further noting $R_N = \Phi_N\Phi_N^T$ yields

$$\begin{aligned}\hat{\theta}_N^R &= (LL^T\Phi_N\Phi_N^T+I)^{-1}LL^T\Phi_N Y_N = \eta L\\ \eta &= L^T\Phi_N(\Phi_N^T LL^T\Phi_N+I)^{-1}Y_N \quad \text{(scalar)}\end{aligned}$$

so the estimate is forced to be parallel to L.

B. Best Regularization For a Known System

The basic result (19) gives a solution to the kernel selection problem. The best that can be achieved by regularization for a known system with impulse response coefficients θ_0 is to let

$$Z^{\text{opt}} = \theta_0\theta_0^T/\sigma^2 \tag{22}$$

This is an interesting insight but cannot be used in practice, since the objective is to find θ_0. It also turns out that the choice may be quite sensitive w.r.t. Z. We shall return, in the next section, to show how this insight could be used in practice.

C. Robustified Choices of Regularization

The optimal choice depends very fundamentally on the given, unknown system. Then a natural question to ask is what is the best choice of Z for a collection of given systems, say, $\theta_0 \in \Theta_\alpha$. This leads to two kinds of strategies:

- Best worst case choice for the given set:

$$Z_\alpha^{\text{opt}} = \arg\min_Z \max_{\theta_0\in\Theta_\alpha} \text{trace}\, M_N(R_N,\theta_0\theta_0^T,Z) \tag{23}$$

- Best average choice for the given set:

$$Z_\alpha^{\text{opt}} = \arg\min_Z E_{\theta_0\in\Theta_\alpha}\text{trace}\, M_N(R_N,\theta_0\theta_0^T,Z) \tag{24}$$

  Even in a frequentist framework we may of course ask for the expected (average) behavior over a set of possible true systems ($\theta_0 \in \Theta_\alpha$).

a) Best Worst Case Choice for White Input: There is an interesting connection between the best worst case choice and the "ideal" choice (19). To illustrate the idea, assume Θ_α is the interior of an ellipsoid of the following form

$$\Theta_\alpha = \{\theta_0 | \theta_0^T\Lambda\theta_0 < \alpha\} \tag{25}$$

where Λ is positive semi-definite and $\alpha > 0$. Also assume sufficient regularity so that we can interchange the minimization and maximization operations in (23). In this case

$$Z = \arg\max_{\theta_0\in\Theta_\alpha}\min_Z \text{trace}\, M_N(R_N,\theta_0\theta_0^T,Z) \tag{26}$$

Now, from (19) we see that the solution to the inner minimization is

$$Z(\theta_0) = \theta_0\theta_0^T/\sigma^2 \tag{27}$$

and hence

$$\min_Z \text{trace}\, M_N(R_N,\theta_0\theta_0^T,Z) \tag{28}$$

$$= \text{trace}((\theta_0\theta_0^T R_N/\sigma^2+I)^{-1}\theta_0\theta_0^T) \tag{29}$$

$$= \frac{\sigma^2\theta_0^T\theta_0}{\theta_0^T R_N\theta_0+\sigma^2} \tag{30}$$

Further assume that the input $u(t)$ is white noise with variance μ so

$$R_N/N \to \mu I_n, \quad N\to\infty \tag{31}$$

Then for sufficiently large N, the outer maximization problem in (26) is equivalent to maximizing

$$\max_{\theta_0} \quad \frac{\sigma^2\theta_0^T\theta_0}{\sigma^2+N\mu\theta_0^T\theta_0} \tag{32}$$

$$\text{subject to} \quad \theta_0^T\Lambda\theta_0 \le \alpha \tag{33}$$

Note that the ratio above is monotonic in $\theta_0^T\theta_0$ so it is a matter of maximizing the norm $\|\theta_0\|$ subject to the constraint (33). This is clearly done by letting θ_0 be proportional to the eigenvector corresponding to the minimum eigenvalue of Λ. Finally, noting (27) yields that the solution to (23) is again of the form (22) with θ_0 corresponding to the smallest eigenvalue of the matrix Λ defined in (25).

b) Best Average Choice: Note that the problem (24) – even without "trace" – can be written

$$Z_\alpha^{\text{opt}} = \arg\min_Z M_N(R_N,Q_\alpha,Z), \qquad Q_\alpha = E_{\theta_0\in\Theta_\alpha}\theta_0\theta_0^T \tag{34}$$

regardless of the shape of the set Θ_α. This follows since $M_N(R_N,\theta_0\theta_0^T,Z)$ is linear in $\theta_0\theta_0^T$. From (19) we know the solution to (34):

$$Z_\alpha^{\text{opt}} = Q_\alpha/\sigma^2 \tag{35}$$

Remark 4.3: The case with averaging over a given subset of systems has an obvious Bayesian interpretation. If we "know" that the system lies in a given set Θ_α where the covariance matrix of θ_0 is Q_α, we can view that as prior information about the parameter. Adding the assumption that the noise $v(t)$ is Gaussian with variance σ^2 and the prior distribution is Gaussian $\theta_0 \in N(0,Q_\alpha)$, gives the posterior density of the parameter, given the observations Y_N as $N(\hat{\theta}_N^R, P_{\text{post}})$. Here, the mean $\hat{\theta}_N^R$ is given by (14) with $Z = Q_\alpha/\sigma^2$ and $P_{\text{post}} = ((\sigma^2 R_N^{-1})^{-1}+Q_\alpha^{-1})^{-1}$.

V. Parameterizations of the "Z" Matrix

To get an idea of how the regularization matrix Z might be parameterized, we first specialize to the case of a diagonal Z-matrix

$$Z = \text{diag}(z_1,...,z_n) \tag{36}$$

and the case of the input $u(t)$ being of white noise, so that for sufficiently large N, according to (31), the n diagonal elements of $M_N(\hat{\theta}_N^R)$ in (16) are readily found to be

$$E(\hat{g}_k^R - g_k^0)^2 = \frac{\sigma^2 N\mu + (g_k^0/z_k)^2}{(N\mu + 1/z_k)^2}, \qquad k = 1,\dots,n \quad (37)$$

For each k, this is minimized by

$$z_k = (g_k^0)^2/\sigma^2 \quad (38)$$

which is well in line with the optimal choice (19). Of course, the true impulse response is not known, but this is a case where we could have some idea about how to parameterize the regularization matrix Z. Assume that the unknown linear stable system has all poles inside a circle with radius $\sqrt{\lambda}$. Then there exists $c > 0$ such that

$$(g_k^0)^2 \le c\lambda^k, \qquad k = 1,\dots,n \quad (39)$$

Therefore we have a natural parametrization of a diagonal regularization matrix

$$Z_\alpha^{DI} = \mathrm{diag}(c\lambda,\dots,c\lambda^n), \qquad \alpha = [c,\lambda] \quad (40)$$

The parameter α is often called a *hyper-parameter* and may not be known, but could be estimated in some ways. We will return to that in the next section.

Thinking that Z in some way should mimic the optimal choice $Z = \theta_0\theta_0^T/\sigma^2$, then λ in the diagonal case presented above captures the decay of the impulse response. We may also try to encapsulate the smoothness of the impulse response. The off-diagonal elements in $\theta_0\theta_0^T$ describe the "correlation" between different parts of the true impulse response. Picking a parameter ρ to describe this smoothness we obtain a matrix, with k,j-element

$$Z_\alpha^{DC}(k,j) = c\rho^{|k-j|}\lambda^{(k+j)/2}, \qquad \alpha = [c,\lambda,\rho] \quad (41)$$

Here $|\rho| \le 1$ and $\rho \approx 1$ means that neighboring values of g_k^0 are very close, while $\rho < 0$ means that neighboring values of g_k^0 tend to have opposite signs.

Remark 5.1: Notice that these assumptions on decay (λ) and smoothness (ρ) of the impulse response coefficients can be given corresponding interpretations about the frequency response of the system, cf the discussion in [6].

We may also link the exponential decay to the correlation (somewhat *ad hoc*) by $\rho = \sqrt{\lambda}$ to obtain the regularization matrix

$$Z_\alpha^{TC}(k,j) = c\min(\lambda^j,\lambda^k), \qquad \alpha = [c,\lambda] \quad (42)$$

and by $\rho = -\sqrt{\lambda}$ to obtain the regularization matrix

$$Z_\alpha^{HF}(k,j) = c(-1)^{k-j}\min(\lambda^j,\lambda^k), \qquad \alpha = [c,\lambda] \quad (43)$$

Remark 5.2: It is interesting to see that the two regularization matrices $Z^{TC}(\alpha)$ and $Z^{HF}(\alpha)$ can also be introduced in a "stochastic" argument in Part I of the companion papers [4]:

- Z^{TC} corresponds to the 1st order stable spline K_1 in eq (10) of [4]. ($\lambda \sim e^{-\beta_1}$)
- Z^{HF} corresponds to high frequency stable spline K_3 in eq (14) of [4]. ($\lambda \sim e^{-\beta_3}$)

In the numerical illustration section, we will also test the so-called 2nd order stable spline kernel [2]:

$$Z_\alpha^{SS}(k,j) = \begin{cases} c\frac{\lambda^{2k}}{2}(\lambda^j - \frac{\lambda^k}{3}), k \ge j \\ c\frac{\lambda^{2j}}{2}(\lambda^k - \frac{\lambda^j}{3}), k < j \end{cases}, \qquad \alpha = [c,\lambda] \quad (44)$$

that corresponds to K_2 in eq (11) of [4]. ($\lambda \sim e^{-\beta_2}$). The scaling factor c in (40) to (44) corresponds to λ_l in (4) of [2].

VI. Estimation of the Hyper-parameter

Among a large number of possible parameterizations of the regularization matrix we have now singled out the particular ones, $Z_\alpha^{DI}, Z_\alpha^{DC}, Z_\alpha^{TC}$ and Z_α^{HF}, based on ideas to mimic the behavior of the optimal (but inaccessible) one, $Z^{\mathrm{opt}} = \theta_0\theta_0^T/\sigma^2$.

They all contain "hyper-parameter" α reflecting assumed decay and smoothness of the unknown impulse response. In a given estimation situation, the parameter α needs to be found, guessed or estimated.

There are several possibilities to do that, for example,

- *Explicitly Minimizing the MSE*
- *Empirical Bayes Method*

A. Explicitly Minimizing the MSE

For a known impulse response θ_0, known variance σ^2 and known input R_N we can compute the MSE (8) for a given regularization matrix Z_α using (16) by

$$f(\alpha,\theta_0,\sigma^2,R_N) = \mathrm{trace}\big((R_N + Z_\alpha^{-1})^{-1} \times (\sigma^2 R_N + Z_\alpha^{-1}\theta_0\theta_0^T + Z_\alpha^{-T})(R_N + Z_\alpha^{-1})^{-1}\big) \quad (45)$$

and then estimate the hyper-parameter α by

$$\hat{\alpha} = \arg\min_\alpha f(\alpha,\theta_0,\sigma^2,R_N) \quad (46)$$

Remark 6.1: A problem is of course that the system θ_0 and the variance σ^2 are not known, but a preliminary estimate $\hat{\theta}_N$ and $\hat{\sigma}^2$ could first be obtained and then the hyper-parameter is found by

$$\hat{\alpha} = \arg\min_\alpha f(\alpha,\hat{\theta}_N,\hat{\sigma}^2,R_N) \quad (47)$$

B. Empirical Bayes Method

A given regularization matrix Z_α can, according to (35), be given a Bayesian interpretation. Assume $\theta \sim N(0,Q_\alpha)$. Then, under the Gaussian assumption of the noise $v(t)$, we find from (11) that

$$Y_N \in N(0,\Sigma_\alpha), \quad \Sigma_\alpha = \sigma^2 I_n + \Phi_N^T Q_\alpha \Phi_N \quad (48)$$

and from the observation of Y_N we can estimate α with the maximum likelihood method:

$$\hat{\alpha} = \arg\min_\alpha Y_N^T \Sigma_\alpha^{-1} Y_N + \log\det\Sigma_\alpha \quad (49)$$

With a known $\hat{\alpha}$, we can compute the corresponding MSE $f(\hat{\alpha},\theta_0,\sigma^2,R_N)$ according to (45), i.e., (17). Recall that in this case the MSE expression is only valid asymptotically as mentioned in Remark 3.1.

VII. Numerical Illustrations

We use the data bank of systems and data sets in Part I of the companion papers [4] for simulations. Like [4], there are four experiments. We refer to [4] for details of the data bank and the four experiments. Here, we use FIR model order of 150 for the 2nd experiment, and 100 for the remaining three experiments.

A. Measure of fit

The quality of an estimated FIR model

$$G(q,\hat{\theta}_N)=\sum_{k=1}^{n}\hat{g}_k q^{-k},\quad \hat{\theta}_N=\begin{bmatrix}\hat{g}_1 & \hat{g}_2 & \dots & \hat{g}_n\end{bmatrix}^T \tag{50}$$

is evaluated according to the MSE in (8). With the knowledge of R_N, $\hat{\sigma}^2$, and the true impulse response θ_0, we can calculate the MSE according to (17) for a given regularization matrix Z. To have a measure that does not depend on the actual size of the impulse response, we will use the following normalized measure of fit (in line with the `compare` command in the System Identification toolbox):

$$\text{fit}=100\left(1-\sqrt{\frac{\sum_{k=1}^{n}E(g_k^0-\hat{g}_k)^2}{\sum_{k=1}^{n}(g_k^0-\bar{g})^2}}\right),\quad \bar{g}=\frac{1}{n}\sum_{k=1}^{n}g_k^0 \tag{51}$$

where the term $\sum_{k=1}^{n}E(g_k^0-\hat{g}_k)^2$ is nothing but (17) for the estimated FIR model (50). The fit (51) is calculated for each estimated FIR model. Each figure in the tables below is an average of the fits, for a particular regularization matrix, over the data bank of data sets.

Remark 7.1: It should be noted that the model fit measures used in [4] (eq (21)) are computed based on the estimated FIR models (50). However, we do not need to know estimated FIR models (50) to compute (51). The expression (17) only involves the true impulse response θ_0, R_N, σ^2, and the regularization matrix Z. The quantities θ_0 and R_N are known from the data bank and σ^2 is estimated from the sample variance of the estimated FIR model of order 150 or 100 using the LS method. The regularization matrix Z_α defined is computed for each system as described in Section V.

B. Sketch of the simulation

For each data set in one of the four experiments, we compute model fit measures for the following four cases:

1) LS method (without regularization).
2) Optimal regularization for completely free regularization matrix. That is, the optimal regularization matrix (22) is applied to get the MSE (17).
3) Optimal regularization within each of the five regularization matrices defined in Section V, using the method of minimizing the MSE. That is, solving (46) for the hyper-parameter α, and apply the resulting one to get the MSE (17). We use (46) rather than (47) to find the theoretical best fit, not depending on a particular preliminary estimate.
4) Similar to 3) but the hyper-parameter α is solved using the Empirical Bayes method (49).

Note that 3) and 4) require global optimization of a non-convex criterion, and we cannot of course guarantee that we always reach that optimum.

C. Simulation results

1) Cases 1) and 2) in Section VII-B: The results are shown in the table below.

Exp. No.	LS without Reg.	Opt. Reg. with (22)
Exp. 1	55.6	97.1
Exp. 2	70.8	98.1
Exp. 3	-3.4×10^6	96.5
Exp. 4	-2.6×10^6	94.7

The very bad fits in the un-regularized LS case in Exp. 3 and 4, are due to the low pass inputs used in those experiments. The matrix R_N is the $n\times n$ covariance matrix of the input, which for $n=100$ has a condition number of $4\cdot10^{12}$. (Essentially the ratio between the largest and the smallest value of the spectrum of the input for this large n). For the un-regularized case the covariance and MSE matrix is essentially R_N^{-1} which explains why regularization is quite necessary.

2) Case 3) in Section VII-B: The five regularization matrices defined from (40) to (44) are tested and correspond to the first five columns of the table below. Moreover, "BEST" denotes the result for the model that has the smallest MSE over all the regularization matrices. The results are shown in the table below.

Exp. No.	DI	SS	HF	TC	DC	BEST
Exp. 1	83.2	82.2	84.3	83.7	86.4	86.7
Exp. 2	88.5	88.5	89.2	89.2	90.7	90.9
Exp. 3	59.3	60.9	48.8	61.4	61.6	63.2
Exp. 4	69.9	89.7	37.7	87.5	89.3	89.8

3) Case 4) in Section VII-B: Similar to the table above, the results using the empirical Bayes method for different regularization matrices are shown in the table below. Here, "BEST" denotes the result for the model that has the largest marginal likelihood over all the regularization matrices. These figures correspond to the analogous results in Fig. 3 of the companion paper [4].

Exp. No.	DI	SS	HF	TC	DC	BEST
Exp. 1	81.9	77.9	82.9	81.9	85.1	85.8
Exp. 2	88.1	85.6	88.5	88.4	90.2	90.5
Exp. 3	52.3	52.8	11.5	54.5	55.2	58.1
Exp. 4	65.0	88.0	-26.2	85.8	87.0	88.1

Recall that each figure in the second and third tables is an average for the 1000 fits obtained for the different systems in the experiments. It is of course interesting to study the distribution of the fits over the different individual data sets. It can be seen from the box plots in Figures 1 and 2 that the method of minimizing the MSE is more robust than the empirical Bayes method, while the figures in the second and third tables look similar.

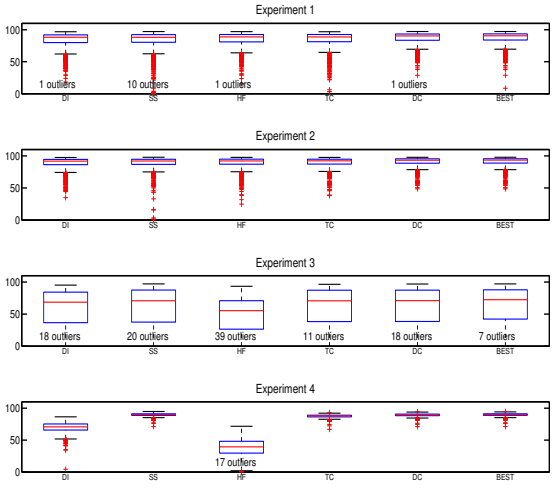

Fig. 1. Box-plots of the fits for the method of minimizing the MSE (46). The plots show left to right the results for the kernels DI,SS, HF, TC, DC and the best choice. Each box plot shows the fit for all the 1000 systems in the corresponding experiment. Fits smaller than 0 (called outliers) are not shown, but the number of such outliers are indicated.

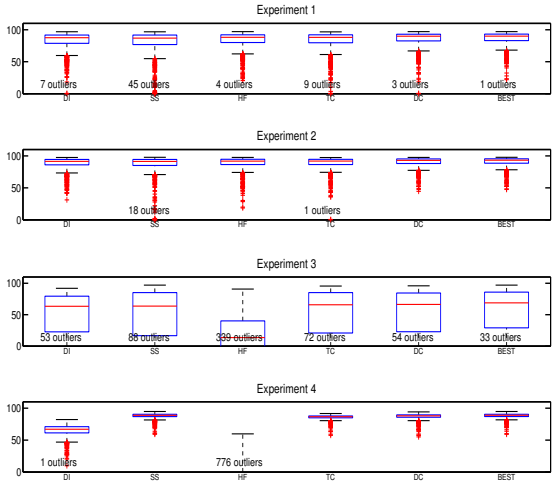

Fig. 2. Box-plots of the fits for the Empirical Bayes method (49). Same legend as in Figure 1.

D. The findings

Not surprisingly, the FIR model obtained using the ideal regularized LS with the optimal regularization matrix (22) works very well for all 4 experiments. The figures reported in the first table are actually the theoretical upper bounds that can be achieved for the FIR model obtained using the regularized LS estimate (14).

The second table shows the theoretical limits for what can be achieved with regularization confined to the particular matrix structures in Section V. It reveals that the constraint in choice of Z causes the fit to drop by 5 to 30 %. Still, quite good fits are obtainable by such regularized LS FIR modeling for a large variatey of systems. The data and systems in Experiment 3 are more difficult.

It is quite remarkable that the empirical Bayes method achieves fits that most of the time are just a few percent units below the theoretical best fit for the kernel in question.

VIII. Conclusions

We have in this paper studied the choice of kernels, or regularization matrices, from a classical, frequentist, perspective. We have explored the boundaries for what can be achieved at all with such kernel methods, if no constraints are placed on the structure of the kernel (see section VII-C.1) showing that very good fits can be achieved for all the systems.

With the constraints of the kernel imposed by the different structures in Section V, the theoretically achievable fit becomes worse, especially for Experiment 3, (see section VII-C.2).

Then the question arises how well these theoretical performance limits can be achieved by algorithms that do not use knowledge of the true system. The empirical Bayes method does very well in that, especially for the kernels SS and DC.

It is actually quite thought-provoking that the empirical Bayes method, which is based on ML estimation of the hyper-parameters comes so close to the theoretical optimal performance for the corresponding regularization kernels. The links between the optimization problems (46) and (49) should be studied more closely.

IX. Acknowledgment

The work was supported by the Swedish Research Council, VR, within the Linnaeus center CADICS. It has also been supported by the European Research Council under contract 267381. The authors express their sincere thanks to Gianluigi Pillenetto for helpful discussions and for making his data bank available to us.

References

[1] L. Ljung. *System Identification - Theory for the User*. Prentice-Hall, Upper Saddle River, N.J., 2nd edition, 1999.

[2] G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, January 2010.

[3] G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 47(2):291–305, February 2011.

[4] G. Pillonetto and G. De Nicolao. Kernel selection in linear system identification. part i: A gaussian process perspective. In *Proc. 50th IEEE Conference on Decision and Control and European Control Conference*, Orlando, Florida.

[5] T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularizations and gaussian processes - revisited. *Automatica*, provisionally accepted, 2011. (An abridged version is to appear in the Proceedings of the 18th IFAC World Congress, Milano, Italy, 2011.)

[6] G. C. Goodwin, J. H. Braslavsky, and M. M. Seron. Non-stationary stochastic embedding for transfer function estimation. *Automatica*, 38:47–62, 2002.