Preprints of the 18th IFAC World Congress
Milano (Italy) August 28 - September 2, 2011

On the Estimation of Transfer Functions, Regularizations and Gaussian Processes - Revisited

Tianshi Chen, **Henrik Ohlsson**, **Lennart Ljung**

Department of Electrical Engineering, Linköping University, Linköping, Sweden (email: {tschen,ohlsson,ljung}@isy.liu.se)

Abstract: Intrigued by some recent results on impulse response estimation by kernel and nonparametric techniques, we revisit the old problem of transfer function estimation from input-output measurements. We formulate a classical regularization approach, focused on finite impulse response (FIR) models, and find that regularization is necessary to cope with the high variance problem. This basic, regularized least squares approach is then a focal point for interpreting other techniques, like Bayesian inference and Gaussian process regression.

Keywords: System identification; transfer function estimation; regularization.

1. INTRODUCTION

Estimation of the transfer function, or impulse response, of a linear system is a problem that we feel that we have known "everything about" for at least a quarter of a century, e.g. Ljung [1985], based on well established theory and algorithms in statistics and the system identification community. Nevertheless, papers on the problem are still appearing. A recent, very inspiring, and thought provoking, contribution is Pillonetto and Nicolao [2010a] (see also the follow-up, Pillonetto et al. [2011]), which shows rather remarkable results based on Gaussian processes and spline kernels. That has prompted the current wish to revisit the transfer function estimation problem from scratch.

Problem Formulation Consider a single-input–single-output linear stable system

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

Here q is the shift operator, $qu(t) = u(t+1)$, $v(t)$ is additive 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 form the *impulse response* of the system. The corresponding frequency function is defined as

$$G_0(e^{i\omega}) = \sum_{k=1}^{\infty} g_k^0 e^{-i\omega k} \tag{3}$$

Given the input-output data $Z^N = \{u(t), y(t), t = 1, \ldots, N\}$, the goal is to find an estimate $\hat{G}_N(e^{i\omega})$ of $G_0(e^{i\omega})$ that is as good as possible. A related goal is to assess and quantify the error in the estimate.

The traditional way is to postulate a finite-dimensional parameterization

$$G(q, \theta) \tag{4}$$

in terms of θ and then estimate θ in some suitable way and deliver the estimate $\hat{G}_N(e^{i\omega}) = G(e^{i\omega}, \hat{\theta}_N)$. Many such parameterizations have been suggested and tested in the literature, e.g. Ljung [1999]. A distinct difficulty is to determine the "size" of the parameter vector θ and to assess the error that stems from $G_0(e^{i\omega})$ being outside the set of functions that is covered within the parameterization. Partly for that reason, alternative approaches based on other ideas, like Gaussian process regression, and non-parametric descriptions of the function $G_0(e^{i\omega})$ (or the impulse response) have recently been suggested, e.g. Pillonetto and Nicolao [2010a], Pillonetto et al. [2011]. Related methods for assessing the quality of $\hat{G}_N(e^{i\omega})$ have been discussed in the 90's and early 2000's, Goodwin et al. [1992], Gustafsson and Hjalmarsson [1995], Goodwin et al. [2002] in connection with bias quantification.

The Question Revisited Suppose we are given a batch of input-output data. We have no information about the data, except that it is collected from a linear stable system with additive noise. The task is to *estimate, as well as possible, the impulse response of the unknown system*. The typical, standard answer to this question is to *apply a prediction error/maximum likelihood (PEM/ML) method to various model structures (4) and use model order/model selection techniques to pick the best model order/structure, and finally compute the impulse response for the model of best order/structure.*

We shall revisit this question with an emphasis on high order regularized FIR (finite impulse response) models, that are simple, safe and robust ways of building linear models, directly focusing on the impulse response. This basic, regularized least squares approach is then a focal point for interpreting other techniques, like Bayesian inference and Gaussian process regression.

2. A DATA-BANK OF TEST DATA

To test different techniques we generated a data-bank of 5000 systems and data sets. They should be representative of real-life data sets, in that the underlying system is not of low order

⋆ The work was supported by the Foundation for Strategic Research, SSF, under the center MOVIII and 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 MATLAB code available to us.

(but could allow good low order approximations) and should correspond to different signal-to-noise ratios (SNR). We have done as follows:

- A number of 30th order random SISO continuous-time systems were generated using the command `rss` in MATLAB.
- These continuous-time systems were sampled at 3 times the bandwidth to yield the discrete-time systems using the following commands in MATLAB
  ```
  bw=bandwidth(m)
  f = bw*3*2*pi
  md=c2d(m,1/f,'zoh')
  ```
  where `m` is the continuous-time system and `md` is the corresponding discrete-time system.
- These discrete-time systems were split into 2500 "fast" systems `S1` that have all their poles inside a circle with radius 0.95 and 2500 "slow" systems `S2` which have at least one pole outside the circle with radius 0.95 (but inside the unit circle).
- The 5000 systems were simulated with an input which was white Gaussian noise with unit variance, and output additive white Gaussian noise with different variances:
  - low SNR: SNR=1. The additive output noise has the same variance as the noise-free output. The number of data in these records is 375.
  - high SNR: SNR=10. The additive output noise has a variance which is a tenth of the variance of the noise-free output. The number of data in these records is 500.

This gives four collections of data sets.

- S1D1: Fast systems with high SNR.
- S2D1: Slow systems with high SNR.
- S1D2: Fast systems with low SNR.
- S2D2: Slow systems with low SNR.

All these data sets are accessible from
`http://www.rt.isy.liu.se/~tschen/research/regul_fir/systems_tested/`

To evaluate the various methods the estimates of the impulse response coefficients $\hat{g}_k$ were compared to the true ones by the measure

$$W = 100\left(1 - \left[\frac{\sum_{k=1}^{n}|g_k^0 - \hat{g}_k|^2}{\sum_{k=1}^{n}|g_k^0 - \bar{g}^0|^2}\right]^{1/2}\right), \quad \bar{g}^0 = \frac{1}{n}\sum_{k=1}^{n} g_k^0 \quad (5)$$

where n is the order of an FIR model and its definition will become clear shortly. The W in (5) corresponds to the "fit" in the `compare` command in the System Identification Toolbox, Ljung [2007]. Note that $W = 100$ means a perfect fit between the true impulse response and the corresponding estimate for the first n coefficients. Each data set gives rise to a particular value of W, and in the tables below we give the average of W over all the sets in a certain collection.

3. A CLASSICAL PERSPECTIVE

In the classical perspective $G_0(e^{i\omega})$ is unknown and estimated from the data. The estimate is a random variable (due to the noise $v(t)$) and the quality can be assessed by the "distance" between the estimate and the true value.

A reasonable measure is the mean square error (MSE)

$$M_N(\omega) = E|\hat{G}_N(e^{i\omega}) - G_0(e^{i\omega})|^2 \quad (6)$$

Here, the expectation E is with respect to the output noise process $v(t)$. Now, the MSE $M_N(\omega)$ is classically split into a bias part

$$B_N(\omega) = E\hat{G}_N(e^{i\omega}) - G_0(e^{i\omega}) \quad (7)$$

and a variance part

$$V_N(\omega) = E|\hat{G}_N(e^{i\omega}) - E\hat{G}_N(e^{i\omega})|^2 \quad (8)$$

so that

$$M_N(\omega) = V_N(\omega) + |B_N(\omega)|^2 \quad (9)$$

3.1 Trading Variance for Bias to Minimize the MSE

In the expression for the MSE $M_N(\omega)$, the bias term $B_N(\omega)$ decreases and the variance term $V_N(\omega)$ increases, when the model becomes more flexible (contains more essential parameters). The MSE $M_N(\omega)$ is then often minimized for a model flexibility that does not give zero bias. In other words, a pragmatic choice of model flexibility allows some bias to reduce variance so that the MSE $M_N(\omega)$ is minimized.

3.2 OE-models

We will not be concerned with noise models in this contribution, so a natural numerator/denominator model is

$$G(q,\theta) = \frac{B(q,\theta)}{F(q,\theta)} \quad (10)$$

The PEM/ML approach to the estimation of (10) would be

$$\hat{\theta}_N^{OE} = \arg\min_{\theta} \sum_{t=1}^{N} |y(t) - G(q,\theta)u(t)|^2 \quad (11)$$

The estimation involves search for the solution of the non-convex problem (11), which may lead to local minima and possibly ill-conditioned calculations. An alternative is to fix the denominator $F(q,\theta)$ to 1 (or any fixed, stable, polynomial) so that a linear regression problem is obtained.

3.3 FIR-models

The simplest approach to estimate $G(q,\theta)$ is to truncate the expansion (2) at a finite number of impulse response coefficients ("FIR" model, corresponding to fixing $F(q,\theta) = 1$ in (10))

$$G(q,\theta) = \sum_{k=1}^{n} g_k q^{-k}, \qquad \theta = [g_1 \ g_2 \ \dots \ g_n]^T \quad (12)$$

where n is the order of the FIR model. The vector θ is then easily estimated by the least squares method. Write the model as

$$y(t) = \varphi^T(t)\theta + v(t), \ \varphi(t) = [u(t-1) \ \dots \ u(t-n)]^T \quad (13a)$$

$$\text{or } Y_N = \Phi_N^T\theta + \Lambda_N \quad (13b)$$

$$\text{where } Y_N = [y(n+1) \ y(n+2) \ \dots \ y(N)]^T \quad (13c)$$

$$\Phi_N = [\varphi(n+1) \ \varphi(n+2) \ \dots \ \varphi(N)] \quad (13d)$$

$$\Lambda_N = [v(n+1) \ v(n+2) \ \dots \ v(N)]^T \quad (13e)$$

The least-squares solution is well known:

$$\hat{\theta}_N^{LS} = \left[\hat{g}_1^{LS} \ \hat{g}_2^{LS} \ \dots \ \hat{g}_n^{LS}\right]^T = \arg\min_{\theta} V_N(\theta) \quad (14a)$$

$$V_N(\theta) = \|Y_N - \Phi_N^T\theta\|^2 = \sum_{t=n+1}^{N} \left(y(t) - \varphi^T(t)\theta\right)^2 \quad (14b)$$

$$\hat{\theta}_N^{LS} = (\Phi_N\Phi_N^T)^{-1}\Phi_N Y_N = R_N^{-1}F_N \quad (14c)$$

$$F_N = \Phi_N Y_N = \sum_{t=n+1}^{N} \varphi(t) y(t) \tag{14d}$$

$$R_N = \Phi_N \Phi_N^T = \sum_{t=n+1}^{N} \varphi(t)\varphi(t)^T \tag{14e}$$

Since $u(-n+1),\dots,u(0)$ are not known, the summation in (14b) starts at $n+1$ to allow $\varphi(t)$ to be formed. This is known as the 'non-windowed' case. As can be seen from (13c), this means that the first n outputs, $y(1), y(2),\dots,y(n)$ in the data set $Z^N = \{u(t), y(t), t = 1,\dots,N\}$ are not used.

How good is the resulting FIR model? Let us assume that

$$Ev(t) = 0, \quad Ev(t)v(s) = \sigma^2 \delta_{t,s} \tag{15}$$

The input $u(t)$ (and thus $\varphi(t)$) is seen as a deterministic variable, and for the conceptual analysis here, for simplicity we will assume that there exists $\mu > 0$ such that

$$\frac{1}{N} R_N \to \mu I_n \quad \text{as } N \to \infty \tag{16}$$

This will hold w.p. 1 if $u(t)$ is chosen as white noise with variance μ but may be true under many other choices of input (PRBS, certain multi-sine input etc). This means that for reasonably large N,

$$\frac{1}{N} R_N \approx \mu I_n \tag{17}$$

Then it is immediate to show that

$$E\hat{\theta}_N^{LS} = \theta_0 = \begin{bmatrix} g_1^0 & g_2^0 & \dots & g_n^0 \end{bmatrix}^T \tag{18}$$

$$E(\hat{\theta}_N^{LS} - \theta_0)(\hat{\theta}_N^{LS} - \theta_0)^T = \sigma^2 R_N^{-1} \approx \frac{\sigma^2}{N\mu} I_n \tag{19}$$

which gives the bias, variance, and MSE, corresponding to (7) to (9), as follows

$$B_N(\omega) = \sum_{k=n+1}^{\infty} g_k^0 e^{i\omega k} \tag{20a}$$

$$V_N(\omega) \approx \frac{n\sigma^2}{N\mu} \tag{20b}$$

$$M_N(\omega) \approx \frac{n\sigma^2}{N\mu} + \left| \sum_{k=n+1}^{\infty} g_k^0 e^{i\omega k} \right|^2 \tag{20c}$$

3.4 Regularization

Still, we see that the variance increases linearly with the FIR model order n so for higher order FIR models it is important to counteract the increasing variance by *regularization*. This is an example of pragmatic bias-variance trade-off, c.f. Section 3.1. Regularization means that we replace the criterion $v_N(\theta)$ in (14) by

$$v_N^R(\theta, D) = \sum_{t=n+1}^{N} \left(y(t) - \varphi^T(t)\theta\right)^2 + \theta^T D \theta \tag{21a}$$

where D is a positive semi-definite $n \times n$ matrix. That changes the estimate to be

$$\hat{\theta}_N^R = \begin{bmatrix} \hat{g}_1^R & \hat{g}_2^R & \dots & \hat{g}_n^R \end{bmatrix}^T = (R_N + D)^{-1} F_N = (R_N + D)^{-1} R_N \hat{\theta}_N^{LS} \tag{21b}$$

How to select D? We have (all expectations are with respect to $v(t)$)

$$MSE(\hat{\theta}_N^R) = (R_N + D)^{-1}(\sigma^2 R_N + D\theta_0\theta_0^T D^T)(R_N + D)^{-1} \tag{22a}$$

where $MSE(\hat{\theta}_N^R)$ is the MSE matrix of $\hat{\theta}_N^R$ with respect to the true impulse response coefficients vector θ_0 in (18).

Suppose that D is diagonal and $D = diag(d_1, d_2, \dots, d_n)$, and (17) is used for R_N. The (k,k)th element of $MSE(\hat{\theta}_N^R)$ satisfies

$$MSE(\hat{g}_k^R) \approx \frac{\sigma^2 \mu N + d_k^2 (g_k^0)^2}{(\mu N + d_k)^2} \tag{23}$$

which is minimized with respect to d_k by $d_k = \sigma^2/(g_k^0)^2$. Therefore this gives a clue how to choose the regularization matrix D: If the system is stable, then there exist positive real numbers c, λ such that the diagonal of D should increase exponentially:

$$d_k = \frac{\sigma^2}{c\lambda^k}, \qquad k = 1,\dots,n \tag{24}$$

Remark 1. Note that the FIR model (12) can be seen as a special case of regularization: If we choose the diagonal regularization $D = diag(d_1, d_2, \dots, d_m)$ with $m > n$ and

$$d_k = \begin{cases} 0 & \text{if} \quad k \le n \\ \infty & \text{if} \quad k > n \end{cases} \tag{25}$$

it is the same as using an FIR model (12).

Remark 2. Regularization as in (21a) is often used in a Tikhonov sense, Tikhonov and Arsenin [1977], where the objective is to make an ill-conditioned problem have better numerical properties. Here, however, the main aspect of regularization is to better deal with the bias-variance trade-off (9).

3.5 Using a Base-line Model

If the impulse response is decaying slowly, a high order FIR model will be required to capture that. It may then be beneficial to incorporate a "base-line model" that can take care of a dominating part of the impulse response. For example, an additive based-line model can be like

$$G(q, \eta, \theta) = G_b(q, \eta) + G_r(q, \theta) \tag{26a}$$

$$\text{with} \qquad G_r(q, \theta) = \sum_{k=1}^{n} g_k q^{-k} \tag{26b}$$

Here $G_b(q, \eta)$ is the base-line model and η is the associated parameter vector and $G_r(q, \theta)$ is a high order FIR model and θ is as defined in (12).

3.6 Cross-validation

Using the classical methods mentioned in Sections 3.2 to 3.4 for optimal MSE means that we must know certain variables (say β), like the best OE model order, the best FIR model order, or the optimal regularization parameters c, λ in (24). The necessary information to compute these are typically not known, which in the classical perspective typically is handled by *cross-validation*:

1) Split the data record into two parts of the same length: an estimation data part and a validation data part.
2) Estimate models $G(q, \hat{\theta}_N)$ using the estimation data for different values of β.
3) Form the error between the measured and the model outputs for these models using the validation data:

$$\varepsilon(t, \beta) = y(t) - G(q, \hat{\theta}_N) u(t) \tag{27}$$

$$W(\beta) = \sum_t |\varepsilon(t, \beta)|^2 \tag{28}$$

and pick the value of β that minimizes $W(\beta)$. The model can then be re-estimated for this β using the whole data record.

3.7 Regularization as Model Merging

A standard way in statistics is to combine two parameter estimates θ_1 and θ_2 with covariance matrices P_1 and P_2 to an estimate

$$\theta = (P_1^{-1} + P_2^{-1})^{-1}(P_1^{-1}\theta_1 + P_2^{-1}\theta_2) \tag{29}$$

That estimate has the smallest variance if the two original ones are unbiased. In that perspective the regularized estimate (21b) can be seen as the combination of the un-regularized estimate $\hat{\theta}_N^{LS}$ and an estimate $\bar{\theta} = [0\ 0\ \dots\ 0]^T$ with variance D^{-1}.

3.8 Numerical Illustration

Let us try these methods on our data bank of data sets as shown in Section 2.

Example 1. (Fixed order OE models). We estimate models (10) of different orders n (same order for $B(q,\theta)$ and $F(q,\theta)$) using the command `m=oe(data,[n,n,1])` in the System Identification Toolbox, Ljung [2007], and compute the average fit (5) for all the models.

The results are shown in the table below. It also contains the fits when the order n for each data set has been chosen by cross-validation (CV) testing orders 5:5:40.

	n=5	n=15	n=25	n=35	n=40	CV
S1D1	86.3	86.4	74.2	54.9	42.6	89.4
S2D1	68.7	71.7	63.1	49.3	42.0	73.2
S1D2	71.9	56.1	34.5	10.2	-1.7	70.8
S2D2	50.8	42.3	20.4	-2.1	-8.5	49.6

Example 2. (Fixed order FIR models). We estimate models (12) of different orders n using the least squares method (14) and compute the average fit (5) for all the models. For fair comparisons we use in all cases the maximum start value of $n = 125$ in (14b).

The results are shown in the table below. It also contains the fits when the order for each data set has been chosen by cross-validation (CV) testing orders 5:10:125.

	$n=5$	$n=35$	$n=65$	$n=95$	$n=125$	CV
S1D1	32.2	83.1	85.8	81.7	76.9	86.1
S2D1	-0.7	47.1	60.0	64.0	65.3	67.4
S1D2	30.8	61.4	46.0	25.9	-0.1	59.6
S2D2	-1.8	30.5	24.2	8.0	-18.2	30.5

Example 3. (FIR-models of order 125 with regularization). We estimate models (12) of order 125 using the regularization method (21) with diagonal D for different values of c and λ in (24), and compute the average fit (5) for all the models. Throughout the simulations in this paper, the variance σ^2 is estimated from the sample variance of the estimated FIR model (12) of order 125 using the least squares method.

The results are shown in the table below. It also contains the fits when c and λ for each data set has been chosen by cross-validation (CV) testing the grid of 9 values, $c = 1, 5, 9$ and $\lambda = 0.5, 0.9, 0.95$.

	$c=1$ $\lambda=0.5$	$c=1$ $\lambda=0.9$	$c=1$ $\lambda=0.95$	$c=9$ $\lambda=0.5$	$c=9$ $\lambda=0.95$	CV
S1D1	51.0	84.8	79.2	58.2	77.5	84.8
S2D1	18.4	67.8	66.8	24.5	65.6	67.1
S1D2	37.4	54.9	36.3	44.7	17.1	55.6
S2D2	6.4	29.5	8.6	12.7	-7.5	23.3

Example 4. (As Example 3 but with base-line model (26)). We estimate models (26) where an additive second order base-line model $G_b(q,\eta)$ is first identified using the command `m=oe(data,[2,2,1])`, then an FIR model (12) of order 125 is estimated using the regularization method as in Example 3.

	$c=1$ $\lambda=0.5$	$c=1$ $\lambda=0.9$	$c=1$ $\lambda=0.95$	$c=9$ $\lambda=0.5$	$c=9$ $\lambda=0.95$	CV
S1D1	74.8	85.4	79.3	78.0	77.5	86.7
S2D1	56.5	72.2	69.6	58.7	68.4	74.1
S1D2	62.2	57.5	37.4	64.3	17.1	66.4
S2D2	42.2	32.6	9.8	42.7	-6.4	45.8

Findings: The "standard" approach (Example 1), works reasonably well. Note that in the simulated data, the "true" order is 30, but this is normally not the best order choice for the OE models. The experiments in Example 2 also show that although the true impulse response is infinite, it is normally not the best choice to use maximum FIR model order. The high variance for such models overrides the low bias. Choosing the FIR model order by cross-validation gives a fit between 30 – 85 %. Using FIR models of order 125 and regularization (21) with diagonal D in (24) (Example 3) does not always improve the fit for all the c, λ tests, and the good affect is largely dependent on their values, so they should be chosen with care. The cross-validation choice of c, λ over the 9 point-grid gives a fit of about the same size as cross-validation over orders. Adding a second order base-line model, (Example 4), is beneficial, mostly so for the slow systems.

4. A BAYESIAN PERSPECTIVE

In the Bayesian view, the parameter to be estimated is itself a random variable, and we seek the posterior distribution of this parameter, given the observations.

In the current setup, we regard the parameter of the nth order FIR model (12), i.e., the impulse response coefficients vector θ as a random variable, say of Gaussian distribution with zero mean and covariance matrix P_n:

$$\theta \sim \mathcal{N}(\theta^{ap}, P_n), \qquad \theta^{ap} = 0 \tag{30}$$

If the input $u(t)$ (and $\varphi(t)$, see (13a)) is known and the noise $v(t)$ is independent Gaussian distributed with

$$v(t) \sim \mathcal{N}(0, \sigma^2) \tag{31}$$

then with

$$Y_N = \Phi_N^T\theta + \Lambda_N \tag{32}$$

Y_N and θ will be jointly Gaussian variables:

$$\begin{bmatrix}\theta\\ Y_N\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}P_n & P_n\Phi_N\\ \Phi_N^T P_n & \Phi_N^T P_n \Phi_N + \sigma^2 I_{N-n}\end{bmatrix}\right) \tag{33}$$

The posterior distribution of θ given Y_N is

$$\theta|Y_N \sim \mathcal{N}(\hat{\theta}_N^{apost}, P_N^{apost}) \tag{34a}$$

$$\hat{\theta}_N^{apost} = P_n\Phi_N(\Phi_N^T P_n\Phi_N + \sigma^2 I_{N-n})^{-1}Y_N \tag{34b}$$

$$P_N^{apost} = P_n - P_n\Phi_N(\Phi_N^T P_n\Phi_N + \sigma^2 I_{N-n})^{-1}\Phi_N^T P_n \tag{34c}$$

We notice that this a posteriori estimate $\hat{\theta}_N^{apost}$ is the same as the regularized estimate $\hat{\theta}_N^R$ if the regularization matrix D is chosen as

$$D = \sigma^2 P_n^{-1} \tag{35}$$

This is just a restatement of the well-known fact that regularization is closely related to prior estimates.

So this gives an insight into how to choose the regularization matrix: Let it reflect the size and correlations of the impulse response coefficients. For the size, it is entirely in line with the choice of diagonal elements (24). If the impulse response is smooth (for example a fast sampled continuous system) it is also natural to let P_n reflect that, by letting the diagonals close to the main diagonal show high correlation. A simple choice is to let the correlation coefficient between g_k and g_j in (12) be $\rho^{|k-j|}$. With diagonal elements of P_n being $c\lambda^k$ as in (24) we then get a covariance matrix P_n whose (k,j)th element is

$$c\rho^{|k-j|}\lambda^{(k+j)/2} \tag{36}$$

where $c,\lambda \geq 0$ and $|\rho| \leq 1$. The estimates that we come up with are thus the same as in the classical, regularized estimate (21b), but the Bayesian perspective has given additional insights into the choice of D.

4.1 Estimating Hyper-parameters

The Bayesian perspective gives one more insight: Suppose that prior knowledge does not give a definite choice of P_n, but it is natural to let it depend on unknown hyper-parameters β, $P_n(\beta)$ (like $\beta = [c\ \lambda]$ in (24)). From (33) we see that

$$Y_N \sim \mathcal{N}(0, \sigma^2 I_{N-n} + \Phi_N^T P_n(\beta)\Phi_N) \tag{37a}$$

so with a classical twist in this Bayesian framework we can form the likelihood function of the observation Y_N given β, and estimate β by the maximum likelihood (ML) method:

$$\hat{\beta} = \arg\min_{\beta} Y_N^T \Sigma(\beta)^{-1} Y_N + \log\det\Sigma(\beta) \tag{37b}$$

where $\Sigma(\beta) = \sigma^2 I_{N-n} + \Phi_N P_n(\beta)\Phi_N^T$. This method of estimating hyper-parameters in the prior distribution is known as the *empirical Bayes* methods.

The noise variance σ^2 used in (37b) and (35) can of course be included among the hyper-parameters, but in the simulations in this paper we used the way as mentioned in Remark 3.

4.2 Numerical Illustration

Let us test, on the data bank of data sets as shown in Section 2, the Bayesian method (34) and (37) with the following prior covariances: the diagonal (24) and the correlation (36)

$$P_{DI}(k,j) = \begin{cases} c\lambda^k & \text{if } k=j \\ 0 & \text{else} \end{cases} \quad \text{('Diagonal')} \tag{38a}$$

$$P_{DC}(k,j) = c\rho^{|k-j|}\lambda^{(k+j)/2} \quad \text{('Diagonal/correlated')} \tag{38b}$$

where the hyper-parameters are $c,\lambda \geq 0$ and $|\rho| \leq 1$. We also test a related prior of (38b) by linking $\rho = \sqrt{\lambda}$:

$$P_{TC}(k,j) = c\min(\lambda^j, \lambda^k) \quad \text{('Tuned/correlated')} \tag{38c}$$

Example 5. (Testing ML estimation of hyper-parameters). We first estimate models (12) of order 125 using the Bayesian method (34) and (37) with the prior covariances (38). Then we estimate models (26) where an additive second order base-line model $G_b(q,\eta)$ is first identified using the command `m=oe(data,[2,2,1])`, then an FIR model (12) of order 125 is estimated using the Bayesian method (34) and (37) again.

The average fit (5) is calculated and the simulation results are shown in table below, where an "e" is appended to the regularization matrix name if a base-line model is used.

	DI	DC	TC	DIe	DCe	TCe
S1D1	86.7	90.8	90.3	88.9	91.2	91.1
S2D1	68.6	78.0	77.8	75.6	81.6	81.6
S1D2	61.8	72.7	72.4	68.9	74.0	74.1
S2D2	33.2	60.7	60.8	50.6	62.2	61.8

Findings: We see that estimating the hyper-parameters for DI and DIe give about the same fit as the CV in Examples 3 and 4. The ML estimates of the hyper-parameters are slightly better though, perhaps since the search is over a continuum of c,λ and not just the 9-point grid, used for CV. It is also clear that allowing and estimating correlation between the impulse response coefficients with DC, and TC gives a clear improvement. It should be noted that the criterion (37b) is not convex, so it requires some care to initialize the search and search for the minimum. This can be illustrated by the fact that TC actually behaves better than DC in some cases, although it is a special case of DC, but with fewer parameters. In all the tests, we initialize $c = exp(5)$, $\rho = 0.5$. Since the optimization problem (37b) is sensitive to the initial value of λ, we solved (37b) twice with two initial values of λ, 1 and 0.5, respectively. The hyper-parameter estimate that gave a larger likelihood $p(Y_N|\beta)$ was chosen as the ultimate hyper-parameter estimate.

5. GAUSSIAN PROCESS REGRESSION TO THE TRANSFER FUNCTION ESTIMATION

Gaussian process regression (GPR) has become a widely spread and very popular method for inference in machine learning, see, e.g. Rasmussen and Williams [2006]. In short, it is about inferring an unknown function $f(x)$ from measurements $y_i, i = 1,2,\ldots,N$ that bear some information about $f(x)$. The argument x can either be a continuous or a discrete variable. The prior information about the function is that it is a Gaussian process, with certain mean and covariance function. This means that the vector $[f(x_1), f(x_2), ..., f(x_n)]$, for any collection of points x_k is a jointly Gaussian random vector, with mean $m(x) = Ef(x)$ and covariances

$$Cov(f(x_k), f(x_j)) = P(x_k, x_j) \tag{39}$$

where $P(x_k,x_j)$ is often called a *kernel*. Often $m(x) \equiv 0$. Typically, the observation y_i is a linear functional of $f(x_i)$, measured in additive Gaussian noise. This causes $[f(x), y_1, \ldots, y_N]$ to be a jointly Gaussian vector, which means that the posterior distributions,

$$p(f(x_1), ..., f(x_n)|y_1, \ldots, y_N) \tag{40}$$

can be calculated by the rules for conditioning jointly Gaussian random variables.

In Pillonetto and Nicolao [2010a] the GPR is applied to estimating the impulse response of a stable linear system. For a sampled model, the impulse response function is given by $g_k^0, k = 1, \ldots, \infty$ in (2). The observation y_i is the measured output in (1) at time $t = i$. Modeling the impulse response function as a Gaussian process means that, for any n,

$$[g_1, \ldots, g_n] \sim \mathcal{N}(0, P_n) \tag{41}$$

where P_n is the $n \times n$ upper left block matrix of the semi-infinite matrix P defined in (39). This is the same situation as in the Bayesian perspective (30)–(34). The Gaussian process estimate of any collections of impulse response coefficients is thus given by (34).

The only thing that remains to be discussed is the choice of prior covariances (41) (or (39)). Of course, the considerations for choosing P_n in (41) and in (30) must be the same, and the relation to the thoughts about the regularization matrix D in (35) still holds. But in GPR several standard choices for (39) exist.

In Pillonetto and Nicolao [2010a] the following kernels/covariance functions are discussed

$$P_{CS}(k,j) = \begin{cases} c\frac{k^2}{2}(j-\frac{k}{3}), k \le j \\ c\frac{j^2}{2}(k-\frac{j}{3}), k > j \end{cases} \quad \text{('Cubic Spline')} \tag{42a}$$

$$P_{SE}(k,j) = ce^{-\frac{(k-j)^2}{2\lambda^2}} \quad \text{('Squared Exponential')} \tag{42b}$$

$$P_{SS}(k,j) = \begin{cases} c\frac{\lambda^{2k}}{2}(\lambda^j - \frac{\lambda^k}{3}), k \le j \\ c\frac{\lambda^{2j}}{2}(\lambda^k - \frac{\lambda^j}{3}), k > j \end{cases} \quad \text{('Stable Spline')} \tag{42c}$$

where the hyper-parameters $c, \lambda \ge 0$. There is also a MATLAB toolbox, Pillonetto and Nicolao [2010b], that implements the GPR, including estimating the hyper-parameters using (37).

Remark 3. As pointed out by the authors of Pillonetto et al. [2011], the prior covariance (38c) (TC) can actually be seen as a stable spline kernel Pillonetto et al. [2011] of order 1.

Let us test the GPR approach with different kernels (42) on the data bank of data sets as shown in Section 2.

Example 6. (D-matrices suggested in the GPR approach). Similar to Example 5, let us estimate the models (12) of order 125 and (26) with the kernels (42).

The average fit (5) is calculated and the simulation results are shown in table below, where an "e" is appended to the kernel name if a base-line model is used.

	CS	SE	SS	CSe	SEe	SSe
S1D1	78.0	80.8	90.3	81.6	84.2	90.4
S2D1	38.8	74.7	71.7	47.9	78.9	81.2
S1D2	16.6	44.4	68.0	60.7	65.7	71.6
S2D2	12.1	48.3	48.2	-44.3	58.6	59.6

Findings: The CS kernel, has difficulties with the slow systems, while the kernel SS shows a performance compatible with DC, DI and TC in Example 5.

Remark 4. For the SS estimate, we used the `SSpline` command in the identification toolbox Pillonetto and Nicolao [2010b] (with `p=125`, `Lab='ny'`, `mv=0`, `mb=1`, `cn=0`, `red=375`, `LP=0` and `LP2=0`). For the remaining estimates, we used our own implementation, which only differs in the estimation of σ^2 and in that two initial values of λ are used in solving (37b) as in Example 5. With our implementation, the four figures for the SS estimate become 90.3, 77.9, 70.1 and 58.5 in order.

Remark 5. It is fair to add that the theory around GPR and its relation to Bayesian estimation is much richer than shown here. The estimation of continuous time impulse responses can be handled in the same framework and there are interesting connections to Reproducing Kernel Hilbert Spaces (RKHS) and spline approximation. Our point here is that the actual resulting impulse response estimate is a regularized FIR estimate (21b) for a certain choices of regularization matrix D. We refer to Pillonetto and Nicolao [2010a] for a more complete account of the theory.

6. CONCLUSION

Let us now sum up the findings about the question posed in the introduction, to estimate the impulse of the unknown system, so that it has the best fit to the true impulse response.

We have tested several algorithms for this on rather large sets of high order systems to assess both approximation and accuracy aspects.

The conventional method is to try the models (10) of different orders using PEM/ML methods, use the cross-validation in Section 3.6 to pick the best model order, and finally use the whole data record to estimate the model (10) with the best model order. This approach was tested in Example 1 where its performance is shown in the column CV.

We have compared the performance of this standard approach with that of the regularization methods based on the kernels/regularization matrices SS, TC and DC as shown in Examples 5 and 6. The results show that the standard approach works reasonably well, but for slow systems with poor signal-to noise ratios a clear improvement is obtained in the average fit for the FIR models with carefully tuned regularization.

Another result of this paper is that the links between "Nonparametric Gaussian Process Regression" and more conventional FIR-modeling with regularized LS-estimates have been exposed.

REFERENCES

G. C. Goodwin, M. Gevers, and B. Ninness. Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Trans. Automatic Control*, 37 (7):913–929, 1992.

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

Fredrik Gustafsson and Håkan Hjalmarsson. Twenty-one ML estimators for model selection. *Automatica*, 31(10):1377–1392, 1995.

L. Ljung. On the estimation of transfer functions. *Automatica*, 21(6):677–696, 1985.

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

L. Ljung. *System Identification Toolbox for use with* MATLAB. *Version 7*. The MathWorks, Inc, Natick, MA, 7th edition, 2007.

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

G. Pillonetto and G. De Nicolao. The stable spline toolbox for system identification. Technical report, University of Padova, Padova, Italy, 2010b.

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.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed Problems*. Winston/Wiley, Washington, D.C., 1977.