Taylor & Francis
Taylor & Francis Group

International Journal of Control

ISSN: 0020-7179 (Print) 1366-5820 (Online) Journal homepage: https://www.tandfonline.com/loi/tcon20

A shift in paradigm for system identification

Lennart Ljung, Tianshi Chen & Biqiang Mu

To cite this article: Lennart Ljung, Tianshi Chen & Biqiang Mu (2019): A shift in paradigm for system identification, International Journal of Control, DOI: 10.1080/00207179.2019.1578407

To link to this article: https://doi.org/10.1080/00207179.2019.1578407

Accepted author version posted online: 03 Feb 2019.
Published online: 12 Feb 2019.

INTERNATIONAL JOURNAL OF CONTROL
https://doi.org/10.1080/00207179.2019.1578407

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

A shift in paradigm for system identification

Lennart Ljung[a], Tianshi Chen[b] and Biqiang Mu[c]

[a]Division of Automatic Control, Department of Electrical Engineering, Linköping University, Linköping, Sweden; [b]School of Science and Engineering and Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, People's Republic of China; [c]Key Laboratory of Systems and Control, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, People's Republic of China

ABSTRACT
System identification is a mature research area with well established paradigms, mostly based on classical statistical methods. Recently, there has been considerable interest in so called kernel-based regularisation methods applied to system identification problem. The recent literature on this is extensive and at times difficult to digest. The purpose of this contribution is to provide an accessible account of the main ideas and results of kernel-based regularisation methods for system identification. The focus is to assess the impact of these new techniques on the field and traditional paradigms.

ARTICLE HISTORY
Received 22 February 2018
Accepted 29 January 2019

KEYWORDS
System identification; linear regressions; bias-variance tradeoff; regularisation; kernel methods

1. Introduction

System identification – to estimate models of dynamical systems from observed input-output data – has been around for more than half a century, with the term coined in Zadeh (1956). Many articles and textbooks on the topic have been written, e.g. Ljung (1999), Söderström and Stoica (1983) and Pintelon and Schoukens (2001). A certain state-of-the art paradigm has emerged, as described in Section 2.

Less than 10 years ago, an alternative approach, called *kernel-based identification*, with inspiration from machine learning, was described by Pillonetto and De Nicolao (2010). This approach gained considerable attention and during the past years the methods have been further developed in more than a hundred scientific papers, many with quite technical and sophisticated contents.

It is the purpose of the present contribution to give an accessible account of the main ideas and results of this approach with a focus on what it means for the paradigm and practical use of System Identification.

2. The traditional system identification setup

In the state-of the art paradigm of system identification, e.g. Ljung (1999), a parameterised model structure

$$\mathcal{M} = \{\mathcal{M}(\theta) | \theta \in D_{\mathcal{M}}\} \tag{1}$$

is defined, where each model $\mathcal{M}(\theta)$ provides a one-step-ahead predictor of the next output: $\hat{y}(t|\theta)$. Here θ is the parameter to be estimated, which ranges over a set of values $D_{\mathcal{M}}$. In this contribution all signals are assumed to take real values. The model structures can be of many different kinds: for linear black box models the state-space model structure of a given order n is a common choice as are several kinds of polynominal models, ARX, ARMAX, Output Error (OE), Box–Jenkins (BJ) models with given orders of the polynomials that define the respective structure, like See, for example, Ljung (1999), Ch 4. The model parameters θ are estimated by Maximum Likelihood (prediction error) methods, (*ML/PEM*), minimising sums of the prediction errors like:

$$\hat{\theta} = \text{argmin}_{\theta \in D_{\mathcal{M}}} \sum_{t=1}^{N} \|y(t) - \hat{y}(t|\theta)\|^2. \tag{2}$$

The structural parameters, i.e. the model orders of polynomials and state are estimated by a separate procedure, like cross validation, or some order criterion like the Akaike Information Criterion, AIC, or the Bayesian order criterion, BIC.

This *state-of the art procedure* is supported by efficient software, and has been very successful, with many important practical applications.

3. Bias and variance

The goal of system identification is to provide as accurate a model as possible. There are two sources of errors in an estimated model $\mathcal{M}(\hat{\theta})$:

Bias The model class $\mathcal{M}$ was not big enough to contain a description of the true system. Then regardless of the amount and quality of the data, a correct model cannot be obtained. The distance between the true system and the best model in the model class is a *bias error* in the model estimate. It is deterministic in nature and does not depend on disturbances in the observed data.

CONTACT Lennart Ljung ✉ lennart.ljung@liu.se
Dedicated to Sasha with friendship and admiration

Variance The measured data is always corrupted by noise and disturbances and they will also affect the estimates. Even if a true model description is available in the model set, there will be a discrepancy between that and the estimate, caused by the disturbances. This error is stochastic in nature and called the *variance error*.

The total discrepancy between the true system and the model estimate will contain both these error components. The resulting *mean square error, MSE* is, conceptually

$$MSE = BIAS^2 + VARIANCE. \tag{3}$$

So, in order to minimise the MSE one must take both components into consideration. The best choice is a trade-off in model flexibility: A large, flexible model set more easily accommodate any system; hence BIAS decreases with increased flexibility. On the other hand, a large model will have many parameters and is more vulnerable to be 'fooled' by disturbances in data. Hence VARIANCE will increase with increased model flexibility.

This BIAS-VARIANCE trade-off in selecting useful model flexibility is at the heart of all estimation problems and in the traditional system identification set-up it is performed when selecting the model order(s), typically by choosing one or several integers.

4. Linear regressions

A very common estimation structure is the *Linear Regression*, where the entity to be explained, the system output y in the case of system identification, is expressed as a linear combination of measured variables φ, plus a noise contribution e:

$$y(t) = \varphi^T(t)\theta + e(t), t = 1, \dots, N. \tag{4}$$

Forming matrices Y_N and Φ_N by stacking $y(t)$ and $\varphi(t), t = 1, \dots, N$, this equation can be rewritten

$$Y_N = \Phi_N\theta + E_N. \tag{5}$$

Such a structure fits the common *ARX-model*

$$\begin{aligned} y(t) + a_1 y(t-1) + \cdots + a_{n_a} y(t-n_a) \\ = b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b) + e(t) \end{aligned} \tag{6}$$

by taking

$$\begin{aligned} \theta &= [a_1 \quad \dots \quad a_{n_a} \quad b_1 \quad \dots \quad b_{n_b}]^{\mathrm{T}} \\ \varphi(t) &= [-y(t-1) \dots - y(t-n_a) \\ &\qquad u(t-1) \dots u(t-n_b)]^{\mathrm{T}}. \end{aligned} \tag{7}$$

The special case of $n_a = 0$ is known as an *FIR* (*Finite Impulse Response*) model.

Many other identification problems can be formulated as a linear regression. In networked systems, for example, a potential node can be modelled by several incoming inputs from d other nodes

$$y(t) = \sum_{i=1}^{d} G_i(q) u_i(t) + e(t). \tag{8}$$

If each of the incoming transfer functions is modelled as a FIR model, it is clear how to form a linear regression (4) analogously to (7). The structure of the network will then be determined by finding out which G_i can be set to zero. Such an application to sparse dynamic network identification is described in Chiuso and Pillonetto (2012), Chen, Andersen, Ljung, Chiuso, and Pillonetto (2014) as well as in Zorzi and Chiuso (2017)

A nonlinear regression model, $y(t) = g(\varphi(t)) + e(t)$ can be dealt with as a linear regression if a function expansion in predetermined basis functions g_k is used:

$$g(\varphi) = \sum_{k=1}^{d} \alpha_k g_k(\varphi). \tag{9}$$

The model will then become a linear regression in the parameters α with $g_k(\varphi(t))$ as the known regressors. See, for example, Chiuso and Pillonetto (2012) and Pan, Yuan, Ljung, Goncalves, and Stan (2017) for such an application.

The great advantage with a linear regression is that the parameters can readily be estimated by the linear least squares method:

$$\begin{aligned} \hat{\theta}_N^{LS} &= \text{argmin}_\theta \|Y_N - \Phi_N\theta\|^2 \\ &= (\Phi_N^T\Phi_N)^{-1}\Phi_N^T Y_N \\ &= R_N^{-1} F_N, \end{aligned} \tag{10}$$

where $R_N = \Phi_N^T\Phi_N,\ F_N = \Phi_N^T Y_N$.

5. Regularisation

It is well known that if indeed the system is described by (4) for a certain 'true' parameter θ_0 and white Gaussian noise e with variance σ^2, then the LS estimate (10) is the Maximum Likelihood estimate, and it will be unbiased,

$$E\hat{\theta}_N^{LS} = \theta_0 \tag{11}$$

with variance

$$E(\hat{\theta}_N^{LS} - \theta_0)(\hat{\theta}_N^{LS} - \theta_0)^{\mathrm{T}} = \sigma^2 R_N^{-1}. \tag{12}$$

This variance is the Cramer–Rao bound, the smallest variance that can be achieved in any unbiased estimator.

In the least squares solution (10) it may happen that the matrix R_N is ill-conditioned. For that reason, *Regularisation* has been suggested, see Tikhonov and Arsenin (1977) for a general treatment of regularisation. This is generally expressed in optimisation form as

$$\begin{aligned} \hat{\theta}_N^R(\Pi) &= \text{argmin}\|Y_N - \Phi_N\theta\|^2 + \sigma^2\theta^T\Pi^{-1}\theta \\ &= [R_N + \sigma^2\Pi^{-1}]^{-1}\Phi_N^T Y_N. \end{aligned} \tag{13}$$

With Π being a suitable positive definite matrix, this will improve the conditioning of the matrix to be inverted. This numerical aspect was an original reason for regularisation.

In a Bayesian setting, where θ is seen as a random vector, one may note that (13) actually defines the *MAP: Maximum a Posteriori* estimate of θ in case there is prior information that the prior distribution of θ is Gaussian with zero mean and covariance matrix Π.

Back to a 'classical' perspective, where θ_0 is unknown but fixed parameter, the statistical properties of the regularised estimate $\hat{\theta}_N^R(\Pi)$ can easily be determined. It will no longer be unbiased:

$$\hat{\theta}^R_{bias} = E\hat{\theta}^R_N(\Pi) - \theta_0 = -(R_N + \sigma^2\Pi^{-1})^{-1}\sigma^2\Pi^{-1}\theta_0 \quad (14)$$

and the variance around the mean will be

$$\sigma^2(R_N + \sigma^2\Pi^{-1})^{-1}R_N(R_N + \sigma^2\Pi^{-1})^{-1}. \quad (15)$$

We realise that if R_N is ill-conditioned, the Cramer–Rao bound, the best possible variance of an unbiased estimate could be extremely large. At the same time, even a small regularisation term, e.g. $\Pi^{-1} = \delta I$ will improve the conditioning of the matrix $R_N + \sigma^2\Pi^{-1}$ and hence the parameter variance considerably.

The small regularisation will also introduce a small bias, but the combined effect on the MSE, see Section 3, could be very beneficial.

This regularisation, with a scaled identity matrix has long been used in statistics under the name of *ridge regression*.

For the regularised estimate, the mean square error (MSE) of the estimate is

$$\begin{aligned} E[(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^{\mathrm{T}}] &= (R_N + \Pi^{-1})^{-1} \\ &\times (\sigma^2 R_N + \Pi^{-1}\theta_0\theta_0^T\Pi^{-1})(R_N + \Pi^{-1})^{-1}. \end{aligned} \quad (16)$$

A rational choice of Π is one that makes this MSE matrix small. How shall we think of good such choices? It is useful to first establish the following Lemma of algebraic nature

Lemma 5.1: *Consider the matrix*

$$M(Q) = (QR + I)^{-1}(QRQ + Z)(RQ + I)^{-1}, \quad (17)$$

where I is the identity matrix with suitable dimension, Q,R and Z are positive semidefinite matrices. Then for all Q

$$M(Q) \geq M(Z), \quad (18)$$

where the inequality is in matrix sense.

The proof consists of straightforward calculations, see, e.g. Chen, Ohlsson, and Ljung (2012).

So, the question what P gives the best MSE of the regularised estimate has a clear answer: *Use*

$$\Pi = \theta_0\theta_0^T. \quad (19)$$

Not surprisingly the best regularisation depends on the unknown system.

Regularisation will thus be a crucial instrument for dealing with the core problem of bias-variance trade-off. Next we illustrate the idea on a simple model example.

6. Example: finite impulse response estimation

Consider the problem to estimate the impulse response of a linear system as an FIR model:

$$y(t) = \sum_{k=0}^{n_b} g(k)u(t-k) + e(t). \quad (20)$$

The choice of order n_b presents a trade-off between bias (large n_b is required to capture slowly decaying impulse responses without too much error) and variance (large n_b gives many parameters to estimate which gives large variance).

Let us illustrate it with a simulated example. We pick a simple second-order butterworth filter as the system:

$$G(z) = \frac{0.02008 + 0.04017z^{-1} + 0.02008z^{-2}}{1 - 1.561z^{-1} + 0.6414z^{-2}}. \quad (21)$$

Its impulse response is shown in Figure 1.

It has decayed to zero after less than 50 time steps. Let us estimate it from data generated by the system. We simulate the system with low-pass filtered white noise as input and add a small white noise output disturbance with variance 0.0025 to the output. One thousand samples are collected. The data is shown in Figure 2.

To determine a good value for n_b we basically have to try a few values and by some validation procedure evaluate which is best. That can be done in several ways, but since we know the true system in this case, we can determine the theoretically best possible value, by trying out all models with $n_b = 1, \ldots, 50$ and find which one has the best fit to the true impulse response. Such a test shows that $n_b = 13$ gives the best error norm (mean square error (MSE) = 0.2522). This estimated impulse response is shown together with the true one in Figure 3.Despite the 1000 data points, with very good signal to noise ratio the estimate is not impressive. The reason is that the low pass input has poor excitation.

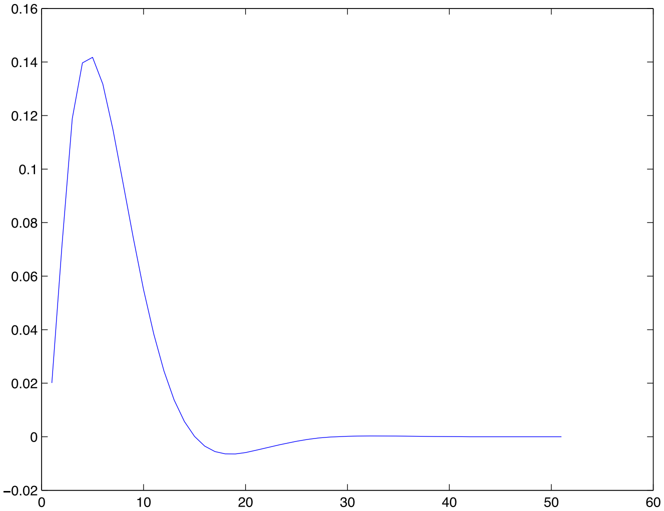

Figure 1. The true impulse response.

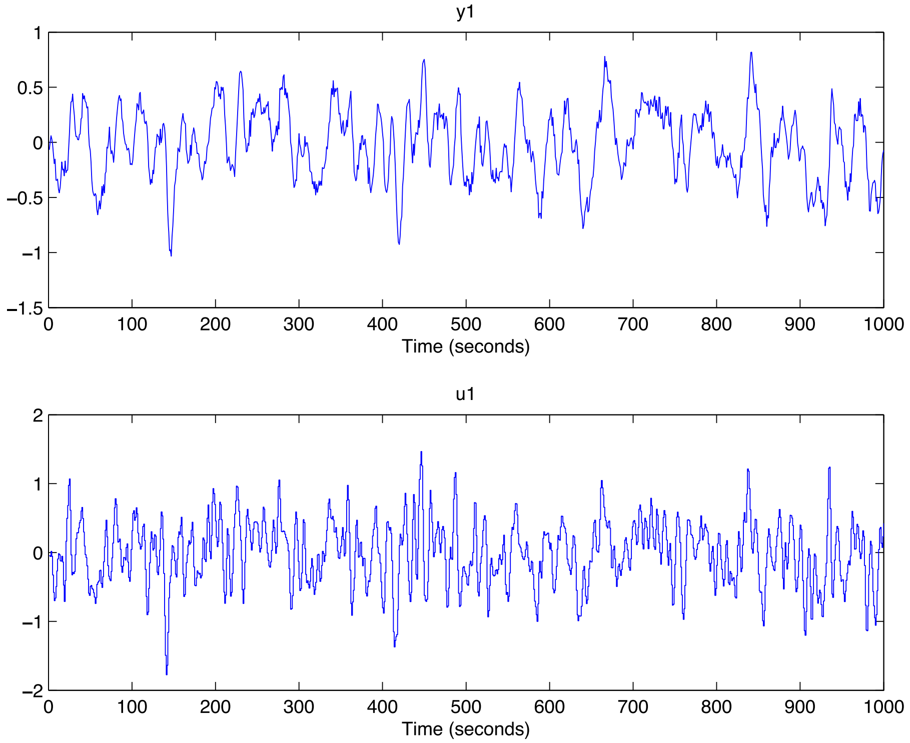

Figure 2. The data used for estimation.

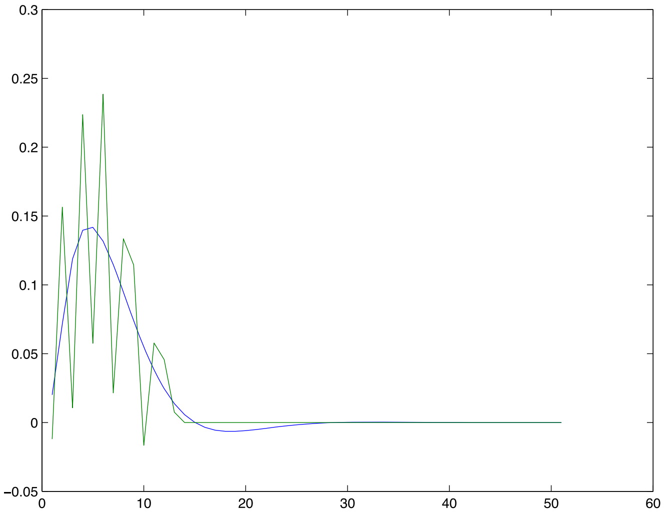

Figure 3. The true impulse response together with the estimate for order $nb = 13$.

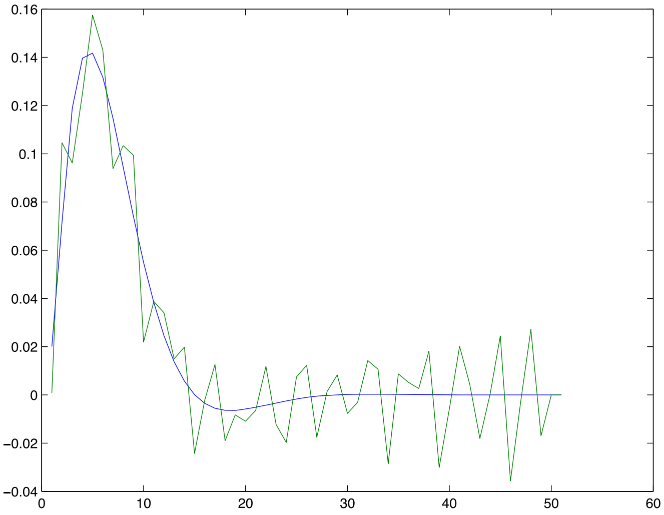

Figure 4. The true impulse response together with the ridge-regularised estimate for order $nb = 50$.

Let us therefore try to reach a good bias-variance trade-off by ridge regression for an FIR model of order 50. That is we use $\Pi = I$ (identity) in (13). The resulting estimate has an MSE error norm of 0.1171 to the true impulse response and is shown in Figure 4.

Clearly even this simple choice of regularisation gives a much better bias–variance tradeoff, than selecting FIR order.

By using the insight that the true impulse response decays to zero and is smooth, we can tailor the choice of Π to the data using the methods to be described in Sections 8 and 9, and obtain the estimate shown in Figure 5 with an MSE norm of 0.0461.

Input design is another important isssue to improve the estimation accuracy of the kernel-based regularisation methods. A two-step procedure is proposed in Mu and Chen (2018), where in the first step, a quadratic transformation of the input is constructed such that the transformed input design problems are convex and their global minima can be found efficiently by applying well-developed convex optimisation software package and in the second step, an expression is derived for the optimal input based on the global minima found in the first step by solving the inverse image of the quadratic transformation.

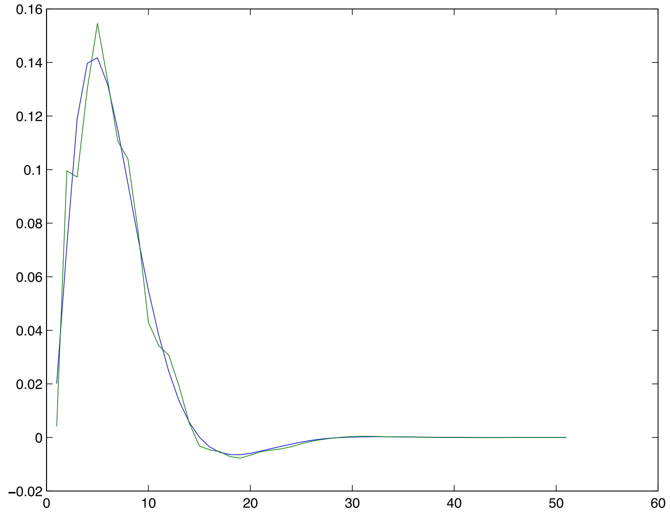

Figure 5. The true impulse response together with the tuned regularised estimate for order *nb* = 50.

7. The two choices

The treatment of regularisation in Section 5 is valid for any linear regression problem. In a sense, the optimal regularisation (for smallest MSE) is known as (19), but this choice depends on the true system and cannot be used in practice. So we have to focus on how to choose Π. It is convenient to split this choice into two parts.

kernel The choice of the structure of Π, that is a parameterisation of the matrix $\Pi(\eta)$. For reasons that will be clear in Section 8, this matrix is also called the *kernel* of the estimation problem. That involves consideration of the actual estimation problem: what is the nature of θ, what properties can be expected.

tuning Once the kernel structure $\Pi(\eta)$ has been selected, the problem is to estimate the *hyperparameters* $\eta \in D_\eta$ using the available observations. This problem as such is a pure estimation problem of statistical nature and does not depend on the actual system.

8. Kernel structure

In this section we assume that the linear regression actually is to estimate the system's impulse response as an FIR model (20). That is, the linear regression as in (6) with $n_a = 0$.That makes the true parameter θ_0 equal to the true impulse response $g^0(k)$.

The kernels for regularisation really have their roots in (continuous) function estimation/approximation techniques, see e.g. Wahba (1999). Using *Reproducing Kernel Hilbert Space*, (*RKHS*) theory, 'basis functions' defined by *splines* appeared as natural and useful elements for function approximation. They define what are called the *kernels* for the function estimation. In our discrete time linear regression formulation, the kernel functions correspond to the Π matrix.

In system identification, the kernel design plays a similar role as the model structure design for ML/PEM and determines the underlying model structure for kernel-based regularisation methods. In the past few years, many efforts have been spent on this issue and many kernels have been invented to embed various types of prior knowledge, e.g. Pillonetto and De Nicolao (2010), Pillonetto, Chiuso, and De Nicolao (2011), Chen et al. (2014, 2016, 2012), Dinuzzo (2015), Carli, Chen, and Ljung (2017), Marconato, Schoukens, and Schoukens (2016), Zorzi and Chiuso (2018), and Pillonetto, Chen, Chiuso, De Nicolao, and Ljung (2016). Among those kernels, the so-called stable spline (SS) kernel Pillonetto and De Nicolao (2010), and diagonal correlated (DC) kernel Pillonetto et al. (2011) are widely recognised and used in the system identification community.

1. In Pillonetto and De Nicolao (2010) the *SS kernel* is derived by applying a wrapping technique to the second-order spline kernel that is used in continuous function estimation. More specifically, the second-order spline kernel $B(s, s')$ is defined on [0,1] × [0,1] and its variance is increasing from 0 to 1 and the corresponding 'stable spline' kernel is defined as $B(\mathrm{e}^{-\beta t}, \mathrm{e}^{-\beta t'})$ with $t, t' \in [0, \infty)$. This leads to the SS kernel, which takes the following form: The regularisation matrix $\Pi(\eta)$ has the k, ℓ element

$$C\left(\lambda^{k+\ell+\max(k,\ell)} - \frac{1}{3}\lambda^{3\max(k,\ell)}\right), \qquad (22)$$

where $\eta = [C, \lambda]$ with $C > 0$ and $0 < \lambda < 1$ are the hyperparameters. Similar idea has also been used to derive the *first-order stable spline kernel*, Pillonetto, Chiuso, and De Nicolao (2010) which takes the following form,

$$C\lambda^{\max(k,\ell)}. \qquad (23)$$

2. In Chen et al. (2012) the *DC kernel* is derived by mimicking the behaviour of the optimal kernel, which is $g^0(t)g^0(t')$ with $g^0(t)$ being the true impulse response of the system to be identified (cf (19)). More specifically, since $g^0(t)$ decays to zero exponentially, it is natural to let the variance of the DC kernel decays with a factor, say λ; since the impulse response is smooth, it is also natural to let the neighbouring impulse response coefficients highly correlated, say with the correlation coefficient ρ. This leads to the DC kernel, which takes the following form,

$$C\lambda^{(k+\ell)/2}\rho^{|k-\ell|}, \qquad (24)$$

where $\eta = [C, \lambda, \rho]$, with $C > 0, 0 < \lambda <, |\rho| \leq 1$ are the hyperparameters. An interesting special case is to take $\rho = \sqrt{\lambda}$. This is called the TC Kernel in Chen et al. (2012), the tuned correlation kernel. Note that in this case the expressions (23) and (24) coincide. This shows the close relationship between the two approaches.This kernel was the one used in Figure 5.

The SS and DC/TC kernels are derived based on heuristic ideas. In practice, we need to develop systematic ways to design kernels to embed various types of prior knowledge. Interestingly, Chen (2018), finds that the SS and DC kernels share some common properties and has developed, based on this finding, two systematic ways of kernel design methods: one is from a machine learning perspective and the other one is from a system theory perspective.

1. Machine Learning perspective: If the impulse response is treated as a function and the prior knowledge is about its decay and varying rate, then the amplitude modulated locally stationary (AMLS) kernel can be designed: a rank-1 kernel and

a stationary kernel are used to account for the decay and varying rate of the impulse response, respectively.

2. System Theory perspective: Suppose that the system's impulse is associated with an LTI system and the prior knowledge is that the LTI system is stable and may be overdamped, underdamped, has multiple distinct time constants or multiple distinct resonant frequencies, etc. Then a simulation induced (SI) kernel can be designed by a nominal model that embeds the given prior knowledge. Also, a multiplicative uncertainty configuration is used to take into account both the nominal model and its uncertainty, and the model is simulated with an impulsive input to get the SI kernel.

9. Hyperparameter tuning

The hyperparameter estimation plays a similar role as the model order selection in ML/PEM and its essence is to determine a suitable model complexity based on the data. As mentioned in the survey of kernel regularisation methods, Pillonetto, Dinuzzo, Chen, De Nicolao, and Ljung (2014), many methods can be used for hyperparameter estimation, such as the cross-validation (CV), empirical Bayes (EB), C_p statistics and Stein's unbiased risk estimator (SURE), etc. Hyperparameter estimation is discussed in, e.g. Chen et al. (2014), Aravkin, Burke, Chiuso, and Pillonetto (2012a, 2012b, 2014), Pillonetto and Chiuso (2015), Mu, Chen, and Ljung (2017), and Mu, Chen, and Ljung (2018).

We now consider the problem to estimate the hyperparameter $\eta \in D_\eta$ in any kernel structure $\Pi(\eta)$ for any linear regression problem (13). As pointed out in Pillonetto and Chiuso (2015) the problem of hyperparameter estimation has certain links with the order selection (tuning complexity) problem. What would be the goal of the estimate? In traditional estimation one seeks estimates, that at least when data sample record tends to infinity converge to the 'true' parameters. What is the 'true' hyperparameter for the current problem?

If we know the true system θ_0, we can for any Π in (13) evaluate the MSE

$$MSE_N(\eta) = E\|\hat{\theta}_N^R(\Pi(\eta)) - \theta_0\|^2. \tag{25}$$

We would like to find an η so that this MSE is minimised asymptotically as $N \to \infty$. A complication is that $MSE_N(\eta)$ will tend to zero for all η. We therefore look at the rate by which the error tends to the Cramer-Rao limit:

$$\begin{aligned} W(P, \Sigma, \theta_0) &= \lim_{N\to\infty} N^2(MSE_N(\eta) - \sigma^2 \mathrm{Tr}(\Sigma^{-1})) \\ &= \sigma^4 \theta_0^T \Pi^{-1}\Sigma^{-2}\Pi^{-1}\theta_0 - 2\sigma^4 \mathrm{Tr}\big(\Sigma^{-1}\Pi^{-1}\Sigma^{-1}\big) \end{aligned} \tag{26}$$

$$\Sigma = \lim_{N\to\infty} \frac{1}{N}\Phi_N^T\Phi_N. \tag{27}$$

Now the 'optimal' hyperparameter will be

$$\eta^*(\Sigma, \theta_0) = \mathrm{argmin}_{\eta\in D_\eta} W(\eta, \Sigma, \theta_0). \tag{28}$$

It depends both on the true system θ_0 and the data Σ.

Remark: Regularisation is essentially a transient phenomenon. As $N \to \infty$ the effect disappears for any η. Still it is meaningful to use the asymptotically defined η^*, since it is a well defined value, and even for large N it gives the best possible MSE.

We would like to optimally estimate η. For the linear regression case with a Gaussian prior for θ we have

$$Y = \Phi\theta + E, \quad \theta \in N(0, \Pi(\eta)). \tag{29}$$

If E is Gaussian with covariance matrix $\sigma^2 I$, Y will also be Gaussian with zero mean and covariance matrix

$$Z(\eta) = \Phi\Pi(\eta)\Phi^T + \sigma^2 I. \tag{30}$$

This means that the probability distribution of Y is known up to η which implies that parameter can be estimated by the maximum likelihood method as

$$\hat{\eta}_N^{EB} = \mathrm{argmin}_{\eta\in D_\eta} Y^T Z(\eta)^{-1} Y + \log\det Z(\eta). \tag{31}$$

This estimate of the prior distribution is known as *Empirical Bayes*. It is a very natural and easy-to-use estimate and is probably the most commonly used one. It normally performs very well, but it is shown in Mu et al. (2018) that it in general does not obey $\hat{\eta}_N^{EB} \to \eta^*$ as $N \to \infty$.

To achieve this goal it is necessary to form another criterion. In Pillonetto and Chiuso (2015) an SURE (Stein's Unbiased Risk Estimator) criterion aiming for minimising the MSE (25) was studied:

$$\begin{aligned} \mathcal{F}_N(\eta) &= \|\hat{\theta}_N^{LS} - \hat{\theta}_N^R(\Pi(\eta))\|^2 + \sigma^2\left(2H^{-1} - (\Phi_N^T\Phi_N)^{-1}\right) \\ H &= \Phi_N^T\Phi_N + \sigma^2\Pi^{-1}(\eta) \end{aligned} \tag{32}$$

with estimate

$$\hat{\eta}_N^S = \mathrm{argmin}_{\eta\in D\eta}\mathcal{F}_N(\eta). \tag{33}$$

It is shown in Mu et al. (2018) that indeed

$$\lim_{N\to\infty} \hat{\eta}_N^S = \eta^*. \tag{34}$$

The estimate is tested in simulations in Mu et al. (2018) and Pillonetto and Chiuso (2015). For well conditioned regressors it behaves well, but if $\Phi_N^T\Phi_N$ is ill-conditioned it suffers from numerical implementation problems.

10. General identification problems

So far we have only dealt with linear regressions (4). They do cover many important estimation problems, but traditional system identification deals with many more models. Linear black-box models,

$$y(t) = G(q)u(t) + H(q)e(t) \tag{35}$$

form a very common class of models, where G and H can be parameterised in several different ways, ARMAX, Ouput-Error, Box-Jenkins or State-space models and estimated by tailored sophisticated identification algorithms.

It is an important and well known fact, Ljung and Wahlberg (1992), that any linear system (35) can be arbitrarily well approximated by an ARX-model (6) with sufficiently large orders,

$$y(t) = \frac{B(q)}{A(q)}u(t) + \frac{1}{A(q)}e(t). \tag{36}$$

This is a linear regression, (7). Thus the regularisation theory can readily be applied to estimate (36) and hence any linear black-box model.

The predictor can be written (cf (6))

$$\hat{y}(t) = -a_1 y(t-1) - \cdots - a_{n_a} y(t-n_a) \tag{37}$$

$$+ b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b) \tag{38}$$

and can be seen as two impulse responses - one from y with the a-parameters and one from u with the b-parameters. Thus the thinking from Section 8 to form the regularisation matrix from impulse response properties can be applied also here.

How does it work? The survey Pillonetto et al. (2014) contains extensive studies with simulated and real data comparing regularised ARX-models with kernels like (24) (applying different kernels to the input and output impulse responses in (7)) with the state-of-the art techniques for ARMAX and state-space models. The result is that the *regularised estimates generally outperform the standard techniques*. The weak spot with the conventional techniques is that it is difficult to find the best model orders. Even an oracle that picks the best possible orders (knowing the true system) in the conventional approach will not do as well as the regularisation. The reason clearly is that the tuning of continuous regularisation parameters – that allows 'non-integer orders' – is a more powerful tool and concept than picking integer order by some selection rule.

11. Conclusions

It is clear that regularisation techniques are a most useful addition to the toolbox of identification methods. The 'automatic' tuning of continuous hyperparameters that determine the flexibility of the used model structure is a powerful and versatile tool compared to the classical rules of selecting model orders. The paradigms in using identification techniques for black box linear systems should definitely allow considering regularisation techniques. In fact, Mathwork's System Identification Toolbox, Ljung, Singh, and Chen (2015) has recently included several of the regularisation techniques described in this contribution. The implementation details can be found in Chen and Ljung (2013).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work has been supported by the National Natural Science Foundation of China under the contracts No. 61603379 and. No. 61773329, the Thousand Youth Talents Plan funded by the central government of China, the Shenzhen Projects Ji-20170189 and Ji-20160207 funded by the Shenzhen Science and Technology Innovation Council, the Presidents grant under contract No. PF. 01.000249 and the Startup grant under contract No. 2014.0003.23 funded by the Chinese University of Hong Kong, Shenzhen, by Shenzhen Fundamental Research Fund under project No. ZDSYS201707251409055, and by a research grant for junior researchers under contract No. 2014-5894 funded by Swedish Research Council.

ORCID

Lennart Ljung http://orcid.org/0000-0003-4881-8955

References

Aravkin, A., Burke, J. V., Chiuso, A., & Pillonetto, G. (2012a). On the mse properties of empirical bayes methods for sparse estimation. In *Proceeding of the IFAC symposium on system identification* (pp. 965–970). Brussels, Belgium.

Aravkin, A., Burke, J. V., Chiuso, A., & Pillonetto, G. (2012b). On the estimation of hyperparameters for empirical bayes estimators: Maximum marginal likelihood vs minimum mse. In *Proceeding of the IFAC symposium on system identification* (pp. 125–130). Brussels, Belgium.

Aravkin, A., Burke, J. V., Chiuso, A., & Pillonetto, G. (2014). Convex vs non-convex estimators for regression and sparse estimation: The mean squared error properties of ard and glasso. *Journal of Machine Learning Research*, *15*(1), 217–252.

Carli, F. P., Chen, T., & Ljung, L. (2017, March). Maximum entropy kernels for system identification. *IEEE Transactions on Automatic Control*, *62*, 1471–1477.

Chen, T. (2018). On kernel design for regularised lti system identification. *Automatica*, *90*, 109–122.

Chen, T., Andersen, M. S., Ljung, L., Chiuso, A., & Pillonetto, G. (2014). System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactions on Automatic Control*, *59*(11), 2933–2945.

Chen, T., Ardeshiri, T., Carli, F. P., Chiuso, A., Ljung, L., & Pillonetto, G. (2016). Maximum entropy properties of discrete-time first-order stable spline kernel. *Automatica*, *66*, 34–38.

Chen, T., & Ljung, L. (2013). Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica*, *50*, 2213–2220.

Chen, T., Ohlsson, H., & Ljung, L. (2012). On the estimation of transfer functions, regularizations and gaussian processes – revisited. *Automatica*, *48*(8), 1525–1535.

Chiuso, A., & Pillonetto, G. (2012, April). A Bayesian approach to sparse dynamic network identification. *Automatica*, *48*, 1107–1111.

Dinuzzo, F. (2015). Kernels for linear time invariant system identification. *SIAM Journal on Control and Optimization*, *53*(5), 3299–3317.

Ljung, L. (1999). *System identification – theory for the user* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Ljung, L., Singh, R., & Chen, T. (2015, Oct). Regularization features in the system identification toolbox. In *IFAC-PapersOnLine*, vol. 48–28, IFAC Symposium on System Identification.

Ljung, L., & Wahlberg, B. (1992). Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra. *Advances in Applied Probability*, *24*, 412–440.

Marconato, A., Schoukens, M., & Schoukens, J. (2016). Filter-based regularisation for impulse response modelling. *IET Control Theory & Applications*, *11*(2), 194–204.

Mu, B., & Chen, T. (2018, November). On input design for regularized LTI system identifiction: Power constrained input. *Automatica*, *97*, 327–338.

Mu, B., Chen, T., & Ljung, L. (2017). Tuning of hyperparameters for fir models – an asymptotic theory. In *Proceedings of the 20th IFAC world congress* (pp. 2818–2823) Toulouse, France.

Mu, B., Chen, T., & Ljung, L. (2018). On asymptotic properties of hyperparameter estimators for kernel-based regularization methods. *Automatica*, *94*, 381–395.

Pan, W., Yuan, Y., Ljung, L., Goncalves, J., & Stan, G.-B. (2017). Identification of nonlinear state-space systems from heterogeneous datasets. *IEEE Transactions Control of Network Systems*, *5*(2), 737–747.

Pillonetto, G., Chen, T., Chiuso, A., De Nicolao, G., & Ljung, L. (2016). Regularized linear system identification using atomic, nuclear and kernel-based norms: The role of the stability constraint. *Automatica*, *69*, 137–149.

Pillonetto, G., & Chiuso, A. (2015). Tuning complexity in regularized kernel-based regression and linear system identification: The robustness of the marginal likelihood estimator. *Automatica*, *58*, 106–117.

Pillonetto, G., Chiuso, A., & De Nicolao, G. (2010). Regularized estimation of sums of exponentials in spaces generated by stable spline kernels. In *Proc. 2010 American Control Conference* (pp. 498–503). Baltimore, MD: Marriott Waterfront.

Pillonetto, G., Chiuso, A., & De Nicolao, G. (2011). Prediction error identification of linear systems: A nonparametric gaussian regression approach. *Automatica*, *47*(2), 291–305.

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

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., & Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, *50*(3), 657–682.

Pintelon, R., & Schoukens, J. (2001). *System identification – a frequency domain approach*. New York, NY: IEEE Press.

Söderström, T., & Stoica, P. (1983). *Instrumental variable methods for system identification*. New York, NY: Springer-Verlag.

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solutions of ill-posed problems*. Washington, DC: Winston/Wiley.

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces, and the randomized GACV. In B. Scholkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods – support vector learning* (pp. 69–88). Cambridge, MA: MIT Press.

Zadeh, L. A. (1956). On the identification problem. *IRE Transactions on Circuit Theory*, *3*, 277–281.

Zorzi, M., & Chiuso, A. (2017). Sparse plus low rank network identification: A nonparametric approach,. *Automatica*, *76*, 355–366.

Zorzi, M., & Chiuso, A. (2018). The harmonic analysis of kernel functions. *Automatica*, *94*, 125–137.