Model Structure Selection – An Update

Håkan Hjalmarsson* and Cristian R. Rojas*

***Abstract*— While the topic has a long history in research, model structure selection is still one of the more challenging problems in system identification. In this tutorial we focus on impulse response modelling, and link classical techniques such as hypothesis testing and information criteria (e.g. AIC) to recent model estimation approaches, including regularisation. We discuss the problem from minimum mean-square error and maximum-likelihood perspectives.**

I. Introduction

Model structure selection is one of the key issues in system identification [1], [2]. Most methods used in system identification originate from statistics, where the topic has a long history [3], [4], [5], [6], [7]. In this tutorial we will focus on three issues:

- The commonality of many of the used methods.
- The relation between regularisation, recently in vogue due to, e.g., sparse estimation, and model structure selection.
- Inherent limitations in model structure selection

Below we give a brief account of these issues. More details will be provided in the tutorial.

II. Estimation of a linear system

In order to be concise, we will assume that the data is generated by

$$Y = \Phi g^o + V \tag{1}$$

where $\Phi \in \mathbb{R}^{N\times n}$ is a known (deterministic) matrix, where $V \sim \mathcal{N}(0, \sigma^2 I)$, with the noise variance σ^2 being unknown, and where $g^o = \begin{bmatrix} g_1^o & \dots & g_n^o \end{bmatrix}^T \in \mathbb{R}^n$ is an unknown parameter vector. The objective is to estimate g^o (and σ^2) from Y and Φ.

The estimation of a stable linear time-invariant single-input single-output dynamical system can be put into this framework by taking n so large that the impulse response beyond time n is negligible. The regression matrix is then given by a Toeplitz matrix built up by the input to the system $\{u(t)\}$ and Y by the output $\{y(t)\}$. More precisely the model is here

$$y(t) = \sum_{k=1}^{n} g_k^o u(t-k) + v(t)$$

so that

$$Y = \begin{bmatrix} y(n+1) \\ \vdots \\ y(N) \end{bmatrix}, \quad \Phi = \begin{bmatrix} u(n) & \dots & u(1) \\ \vdots & \dots & \vdots \\ u(N-1) & \dots & u(N-n) \end{bmatrix}$$

Without any other assumptions, estimating g^o in (1) is the classical linear regression problem. Omitting some parameter independent terms and rescaling the remaining term, the negative log-likelihoood criterion is given by

$$J_{ML}(g) := \|Y - \Phi g\|^2 = (Y - \Phi g)^T (Y - \Phi g) \tag{2}$$

A. Unstructured estimation

The unstructured Maximum-Likelihood (ML) estimate of g^o is defined as

$$\hat{g}_{ML} := \arg\min_g J_{ML}(g) \tag{3}$$

Completing the square gives

$$J_{ML}(g) = J_{MR}(g) + J_{ML}(\hat{g}_{LS}) \tag{4}$$

where

$$\begin{aligned} J_{MR}(g) &:= (g - \hat{g}_{LS})^T R (g - \hat{g}_{LS}) \\ R &:= \Phi^T \Phi \\ \hat{g}_{LS} &:= R^{-1}\Phi^T Y = g^o + R^{-1}\Phi^T V \\ J_{ML}(\hat{g}_{LS}) &= Y^T (I - P_\Phi) Y = V^T (I - P_\Phi) V \end{aligned}$$

where in turn P_Φ is the projection matrix

$$P_\Phi := \Phi R^{-1} \Phi^T$$

From (4) it is immediate that the unstructured ML estimate is given by $\hat{g}_{ML} = \hat{g}_{LS}$, i.e. the least-squares estimate. This estimate is unbiased

$$\mathrm{E}[\hat{g}_{ML}] = \mathrm{E}[g^o + R^{-1}\Phi^T V] = g^o$$

and attains the Cramér-Rao bound for the unstructured case, i.e.

$$\begin{aligned} \mathrm{E}[(\hat{g}_{ML} - g^o)(\hat{g}_{ML} - g^o)^T] &= \mathrm{E}[R^{-1}\Phi^T V V \Phi R^{-1}] \\ &= \sigma^2 R^{-1} \end{aligned}$$

is the smallest possible covariance matrix for an unbiased model given the unstructured model (1).

This work was supported by the European Research Council under the advanced grant LEARN, contract 267381 and by the Swedish Research Council under contract 621-2009-4017.

*ACCESS, School of Electrical Engineering, KTH, SE-100 44 Stockholm, Sweden. (e-mails: {hakan.hjalmarsson, cristian.rojas}@ee.kth.se).

B. Structured estimation

A model structure is a restriction on g^o represented by a set $\mathcal{G}$:

$$g^o \in \mathcal{G} \subset \mathbb{R}^n$$

The (structured) Maximum Likelihood (ML) estimate is then given by

$$\begin{aligned} \hat{g}_{\mathcal{G}} = \arg\min_{g} \; & J_{ML}(g) \\ \text{s.t. } & g \in \mathcal{G} \end{aligned} \tag{5}$$

When $g^o \in \mathcal{G}$, $\hat{g}_{\mathcal{G}}$ will have better accuracy than $\hat{g}_{ML}$.

C. Model structure selection

Consider now that there is a family of candidate model structures

$$\Xi := \{\mathcal{G}(\rho) : \rho \in D_\rho \subset \mathbb{R}^{n_\rho}\}, \tag{6}$$

indexed by the parameter ρ. The model structure selection problem is to select an appropriate structure, i.e. which ρ to use.

D. Relation to sparse estimation

The recent area of sparse estimation is intimately related to model structure selection. To see this, consider as an example the following model structures, corresponding to FIR models of increasing order:

$$\begin{aligned} \mathcal{M}_1: \quad & y_t = g_1 u_{t-1} + e_t, \quad e_t \sim \mathcal{N}(0, \sigma^2) \\ \mathcal{M}_2: \quad & y_t = g_1 u_{t-1} + g_2 u_{t-2} + e_t, \quad e_t \sim \mathcal{N}(0, \sigma^2) \\ & \vdots \end{aligned}$$

As in this example, many model selection problems consider *nested* structures, where$\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \cdots \subseteq \mathcal{M}$. For these problems, we can re-parametrise the structures so that

$$\begin{aligned} \mathcal{M}_1: \; & \tilde{\theta}_1 = (\theta_1, \, 0, \ldots, 0) \in \Theta \\ \mathcal{M}_2: \; & \tilde{\theta}_2 = (\theta_1, \theta_2, \ldots, 0) \in \Theta \\ & \vdots \end{aligned}$$

where Θ is a parameter space associated with the largest model structure into consideration, $\mathcal{M}$. The problem of model structure selection then corresponds to deciding how many trailing zeros $\tilde{\theta}$ should have.

In sparse estimation, the problem is very similar: θ is assumed to be *sparse*, i.e., it is supposed to consist of very few nonzero entries. The goal of sparse estimation is then to determine which entries are zero, and, afterwards, to estimate the value of the non-zero entries. The first step in sparse estimation can be seen, therefore, as one of model selection!

III. Approaches to model structure selection

There are several well-established approaches to model structure selection. A very popular method for model structure selection is *cross-validation*, where the performance of a model is assessed on a validation data set, i.e., a fresh data set not used to identify the parameters [8], [9].

Information criteria try to mimic cross-validation without using a separate data set. The celebrated AIC and BIC criteria [3], [10], [11] belong to this class of methods.

Hypothesis testing can also be used for model structure selection [12]. Here the idea is to pick the simplest structure that cannot be rejected by a statistical test based on the assumptions associated with the structure in question. Most standard model selection criteria can be re-phrased as hypothesis tests [13]. A very common test statistic measures the sample cross-correlation between the inputs and the residuals $Y - \Phi g$ [14]. If the sample cross-correlation has too large magnitude, the model structure is deemed inadequate.

IV. A common format for selection criteria

Returning to (4), we see that the structured estimation problem (5) can be expressed as

$$\begin{aligned} \hat{g}_{\mathcal{G}} = \arg\min_{g} \; & J_{MR}(g) \\ \text{s.t. } & g \in \mathcal{G} \end{aligned} \tag{7}$$

We thus see that the model structure selection problem can be seen as model reduction of the unconstrained (unstructured) ML-estimate $\hat{g}_{ML}$. From this insight it follows that there is a close connection betweeen the different model structure approaches. It turns out that many of them can be cast as tests of the type

$$\text{Reject } \mathcal{G} \text{ if } J_{MR}(\hat{g}_{\mathcal{G}}) > c\,\hat{\sigma}^2 \tag{8}$$

where $\hat{\sigma}^2$ is an estimate of the noise variance σ^2, and where the constant c in (8) depends on the particular test. As we will discuss the particular noise variance estimate has a big impact on the result.

V. Model structure selection and regularisation

The mean-squared error (MSE) of an estimate $\hat{g}$

$$\text{MSE}(\hat{g}) := \text{E}[\|\hat{g} - g^o\|_2^2] \tag{9}$$

can be split into a bias term and a variance term according to

$$\text{MSE}(\hat{g}) := \underbrace{\|\text{E}[\hat{g}] - g^o\|_2^2}_{\text{Bias}^2} + \underbrace{\text{E}[\|\hat{g} - \text{E}[\hat{g}]\|_2^2]}_{\text{Variance}} \tag{10}$$

The MSE of an estimate can be decreased by introducing, on purpose, some bias in such a way that the variance decreases more than the contribution to the MSE from the bias term. Methods aiming to improve the accuracy of the unstructured estimate $\hat{g}_{ML}$ in this way can be seen as alternatives to structured estimation (5). By carefully tuning the bias, it is possible to obtain estimates which are uniformly better than

ML in terms of MSE (over all values of g^o); this is typically called *Stein's phenomenon* [15].

One method to achieve this is regularisation. Here, the ML-criterion (2) is modified to

$$J_P(g, \alpha, \rho_1) := J_{ML}(g) + \alpha P(g, \rho_1) \tag{11}$$

where P is a penalty function that penalises certain properties of g, and where $\alpha \geq 0$ is an appropriately chosen constant, and $\rho_1 \in \mathbb{R}^m$ is a parameter that can be used to shape P.

A classical choice of P is

$$P(g) = \|g\|_2^2 \tag{12}$$

leading to so-called ridge-regression. The penalty function (12) pulls the estimate towards the origin, thereby decreasing its variance (take the extreme case $\alpha = +\infty$ which always yields zero as estimate) at the cost of some bias.

In [16], so called kernel-based methods were adapted to impulse response estimation. In [17] it was shown that this corresponds to regularisation with a weighted ℓ_2 norm $P(g, \rho_1) = \|g\|_W^2 := g^T W(\rho_1) g$, and α and ρ_1 estimated using Empirical Bayes [18]. With $W(\rho_1)$ tailored to the specifics of impulse responses of finite dimensional systems, very impressive results have been achieved.

Here we link regularisation to model structure selection by the observation that there is a one-to-one correspondence between all estimates that can be obtained by minimising (11), and those that can be obtained solving

$$\begin{aligned} &\min_g J_{ML}(g) \\ &\text{s.t. } P(g, \rho_1) \leq \rho_2 \end{aligned} \tag{13}$$

for $\rho_2 \in [0, \infty)$. Now (13) corresponds to (5) with

$$\mathcal{G} = \mathcal{G}(\rho) := \{g : \ P(g, \rho_1) \leq \rho_2\} \tag{14}$$

where $\rho = \begin{bmatrix} \rho_1^T & \rho_2 \end{bmatrix}^T$. Regularisation thus corresponds to model structure selection. We will elaborate further on this connection in the tutorial. For example, taking $P(g) = \|g\|_1$ leads to the LASSO method and promotes sparse estimates, i.e. that the support of g is taken as small as possible. Research in this direction has been intense.

VI. Fundamental limitations in model structure selection

Many model structure selection methods aim at recovering the true underlying structure. For example, with ρ in (6) denoting the model order, it is desirable to recover the true order asymptotically as the number of observations grows. Or, in sparse estimation, where it is known that some, but not which, of the elements of g are zero, it is of interest to recover this sparsity pattern.

One of the major contributions to model structure selection in recent years has been to show that there exist a fundamental limitation so that methods that achieve this have very poor worst case performance [19]. Indeed, due to their asymptotic properties, model selection criteria can be broadly classified into two groups:

- "AIC-type" (e.g., CV, AIC, GCV, SURE). These criteria are based on an estimate of the *prediction error*, and attempt to choose the model structure with best prediction accuracy. Unfortunately, these criteria are typically inconsistent, that is, they tend to choose model structures which are larger than the smallest containing the true system (in case such a structure exists).
- "BIC-type" (*e.g.* BIC, MDL). These criteria attempt to choose the "true" structure, and, as a result, are typically *consistent*. However, the resulting model does not always lead to the best prediction accuracy.

There is an extensive bibliography on both groups of techniques, and for decades researchers tried to combine their virtues of prediction accuracy and consistency. However, recently it has been shown that such goal is impossible to achieve [20], [21]. In essence, it has been established that the use of any estimator together with any consistent model selection criterion gives an MSE which decays slower than $1/N$ for some g^o, i.e., slower than a standard estimator (e.g., ML) without model selection!

It has also been shown that estimation the model structure also imposes limitations in how the user can assess the model error [22]. We elaborate on these issues and their implications for how to select an appropriate model selection criterion [23].

References

[1] T. Söderström, "Identification: Model structure determination," *Systems and Control Encyclopedia*, 1987.

[2] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[3] H. Akaike, "A new look at statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19(6), pp. 716–723, 1974.

[4] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76(2), pp. 297–307, 1989.

[5] A. D. R. McQuarrie and C.-L. Tsai, *Regression and Time Series Model Selection*. World Scientific, 1998.

[6] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd Edition*. Springer-Verlag, 2002.

[7] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*. Cambridge Univ. Press, 2008.

[8] G. Golub, M. Heath, and G. Wahba, "Generalised cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[9] J. Shao, "Linear model selection by cross-validation," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 486–494, Jun. 1993.

[10] S. Konishi and G. Kitagawa, "Generalised information criteria in model selection," *Biometrika*, vol. 83, no. 4, pp. 875–890, 1996.

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

[12] M. Kendall and A. Stuart, *Advanced Theory of Statistics*. London, UK: Griffin, 1961.

[13] T. Söderström, "On model structure testing in system identification," *International Journal of Control*, vol. 26(1), pp. 1–18, 1977.

[14] N. Draper and H. Smith, *Applied Regression Analysis, 2nd ed.* Wiley, New York, 1981.

[15] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, 1956, pp. 197–206.

[16] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, Jan 2010.

[17] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of tranfer functions, regularizations and Gaussian processes - Revisited," *Automatica*, vol. 48, no. 5, pp. 1525–1535, 2012.
[18] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: John Wiley & Sons, 1998.
[19] H. Leeb and B. Pötscher, "Sparse estimators and the oracle property, or the return of Hodges' estimator," *Journal of Econometrics*, vol. 142, pp. 201–211, 2008.
[20] Y. Yang, "Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation," *Biometrika*, vol. 92(4), pp. 937–950, 2005.
[21] H. Leeb and B. M. Pötscher, "Sparse estimators and the oracle property, or the return of hodges' estimator," *Journal of Econometrics*, vol. 142, pp. 201–211, 2008.
[22] H. Leeb and B. Pötscher, "Can one estimate the conditional distribution of post-model-selection estimators?" *Annals of Statistics*, vol. 34, no. 5, pp. 2554–2591, 2006.
[23] J. Shao, "An asymptotic theory for linear model selection," *Statistica Sinica*, vol. 7, no. 2, pp. 221–242, 1997.