CONSTRAINED AND REGULARIZED SYSTEM IDENTIFICATION

Tor A. Johansen *

* *SINTEF Electronics and Cybernetics, Automatic Control Department, N-7034 Trondheim, Norway. Email: Tor.A.Johansen@ecy.sintef.no.*

Abstract: Prior knowledge can be introduced into system identification problems in terms of constraints on the parameter space, or regularizing penalty functions in a prediction error criterion. The contribution of this work is mainly an extention of the well known FPE (Final Prediction Error) statistic to the case when the system identification problem is constrainted and contains a regularization penalty. The FPECR statistic (Final Prediction Error with Constraints and Regularization) is of potential interest as a criterion for selection of both regularization parameters and structural parameters such as order.

Keywords: Regularization, Optimization, Parameter Estimation, Nonlinear Systems.

1. INTRODUCTION

In practical system identification it is often desirable to introduce prior knowledge into the problem, rather than relying completely on the data. If the model structure is assumed to be fixed, there are still several approaches, cf. Figure 1:

(1) Constraints on the parameter space, for example to ensure stability (Tulleken 1993, Johansen 1996*a*), convexity of an optimal control criterion (Foss and Johansen 1997), fulfillment of balance equations and steady-state data (Kramer *et al.* 1992, Thompson and Kramer 1994), frequency-domain data (Eskinat *et al.* 1993, Eskinat 1995), and explicit belief about parameter values (Bai and Sastry 1986). When the model is overparameterized, additional equality constraints are needed to make the problem well-posed and to avoid trivial solutions (Gawthrop *et al.* 1992, De Moor *et al.* 1994, Moons and De Moor 1995).
(2) Regularization, i.e. penalties on non-smooth behavior of the model (Tikhonov and Arsenin 1977, Larsen and Hansen 1994, Johansen 1997), deviation from explicit belief about parameters, and deviation from a prior known model (Kramer *et al.* 1992, Thompson and Kramer 1994, Johansen 1996*a*). Regularization is a general method that improves the robustness and allows identifiation of non-parsimoneous models (Dayal and MacGregor 1996, Sjöberg *et al.* 1994, Sjöberg *et al.* 1993).

Prior knowledge in terms of constraints and penalties can be implemented directly in a prediction error method (PEM) framework, (Johansen 1996*a*), or the penalties can be reformulated into equivalent prior distributions in a Bayesian system identification framework (Peterka 1981, Tulleken 1993, Kárný *et al.* 1995). Relevant optimization methods are discussed in (Kunisch and Sachs 1992).

The main idea is that explicit application of prior knowledge will improve the robustness of the identification algorithm, eventually leading to more accurate or useful parameter estimates, in some sense.

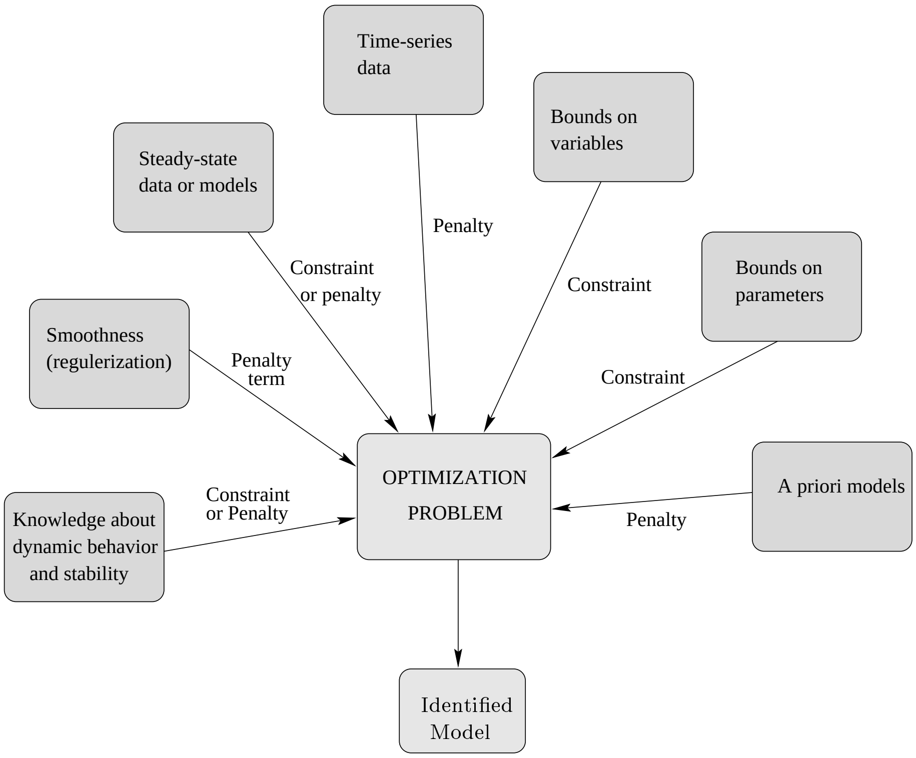

Fig. 1. Prior knowledge in terms of penalties and constraints can be combined with empirical data in an optimization formulation of the identification problem.

The continuation of this paper is as follows: In section 2, we formulate the identication problem with regularization and constraints as an optimization problem, taking the standard PEM as the starting point. The solution to this problem is briefly discussed. The main part of this paper, section 3, describes how the MSE (Mean Squared prediction Error) can be estimated. With the standard PEM, the FPE statistic is a well known estimate that is asymptotically unbiased under some conditions (Akaike 1969, Söderström and Stoica 1988). A generalization of the FPE statistic to cover regularized models was derived in (Larsen and Hansen 1994). Here, this statistic is generalized further to also cover constrained and regularized system identification problems.

2. REGULARIZATION AND CONSTRAINTS

Suppose a model structure, i.e. a set of equations parameterized by a d-dimensional parameter vector θ, is given. The parameter vector can be estimated using a standard prediction error estimator, e.g. (Söderström and Stoica 1988)

$$\hat{\theta}_{PE}(Z^N) = \arg\min_{\theta \in D_c} V_N(\theta; Z^N)$$

$$V_N(\theta; Z^N) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t;\theta)$$

on the basis of a finite data sequence $Z^N = ((u(1), y(1)), (u(2), y(2)), ..., (u(N), y(N)))$, where $\varepsilon(t;\theta) = y(t) - \hat{y}(t; Z^{t-1}, \theta)$, and $y(t)$ and $u(t)$ are the system's scalar outputs and inputs at time t, respectively. The one-step-ahead prediction $\hat{y}(t; Z^{t-1}, \theta)$ is computed by solving the model equations with the parameter vector θ.[1] The parameter set D_c is assumed to be compact, and the predictor is assumed to satisfy the necessary smoothness conditions such that a unique minimum of V_N exists.

If the identifiability of the model is poor, or the data are not sufficiently informative, or the model structure is over-parameterized or fundamentally wrong, the prediction error method may not be robust, giving highly uncertain estimates. In general, some form of prior knowledge can be applied to improve the robustness of the identification problem. An approach to introduce prior knowledge in terms of penalties and constraints to the prediction error method was described in (Johansen 1996a), and briefly reviewed in the introduction. Such penalties are closely related to the method of regularization (Tikhonov and Arsenin 1977) as discussed in (Johansen 1996a, Johansen 1997). Regularization is a general method for improving the robustness of mathematical algorithms by imposing additional regularity constraints on the solution. Mathematically, the problem we are now suggesting to solve has the form

[1] The extension to multi-step-ahead predictors and systems with multiple inputs or outputs is straightforward.

$$V_{N,\gamma}^{\mathrm{REG}}(\theta; Z^N) = V_N(\theta; Z^N) + \gamma\Omega(\theta)$$

subject to

$$G(\theta) = 0,\ K(\theta) \leq 0$$

where G and K are smooth functions defining the equality and inequality constraints, Ω is a stabilizer for the problem, and $\gamma > 0$ is a regularization parameter. [2]

The idea is that the penalty term Ω will attract any uncertain parameters in the model structure towards reasonable regions of the parameter space. Uncertain parameters are characterized by a low sensitivity of the criterion $V_N(\theta; Z^N)$ with respect to perturbations in the corresponding sub-manifold of the parameter space. Hence, the penalty $\gamma\Omega(\theta)$ should contribute significantly to the criterion (related to $V_N(\theta; Z^N)$) when θ is in this sub-manifold. As mentioned in the introduction, the penalties can be selected on the basis of prior knowledge and desired properties such as e.g. smoothness, stability, convexity, additional equations not in the model, and explicit parameter knowledge.

The parameter estimate minimizing the above constrained and regularized prediction error criterion is denoted $\hat{\theta}_{\mathrm{REG},\gamma}$. The asymptotic properties of this estimator in terms of bias and variance is discussed in detail in (Johansen 1997) for the case when there are no constraints. A particular interesting property is that the total parameter error (bias plus variance) can be smaller than the Cramer-Rao lower bound, which is valid only for unbiased estimates such as the PEM. It is expected that if the constraints are based on "correct prior knowlegde", the asymptotic properties are unchanged in the constrained case. If this assumption is violated, additional bias may be introduced, but the variance is decreased.

The estimator properties when there are only constraints (no regularization) is discussed in some generality in (Eskinat 1995).

3. THE FINAL PREDICTION ERROR

We define the MSE (Mean Squared prediction Error) as

$$\mathrm{MSE} = E\left(y(t) - \hat{y}(t; Z^{t-1}, \hat{\theta}(Z^N))\right)^2$$

for some arbitrary estimator $\hat{\theta}(Z^N)$. The variables Z^N, Z^{t-1} and $y(t)$ are viewed as stochastic and E is the expectation with respect to the joint distribution of these variable, so MSE is the ensemble average over all possible identification data sequences of length N, and future data sequences.

The MSE contains information about the expected prediction performance of the identified model. Such information is valuable for a number of purposes, including validation, selection of model structure and order, selection of the regularization parameter γ, and finding the optimal balance between bias and variance, see also (Xin *et al.* 1995). Unfortunately, since the underlying probability measures in system identification problems usually are unknown, the MSE must be estimated empirically. There exists a wide number of methods:

- A separate data sequence (independent of the identification data) can be used to estimate the MSE empirically. The major drawback is that an extra data sequence is needed. The other techniques we will discuss do not require this.
- The method of cross-validation is a widely applied technique (Stone 1974). The idea is to use a subsequence of the data for identification, and the remaining for estimating the MSE. This process is repeated for different subsequences, and the results are averaged. A closely related resampling technique is bootstrapping, e.g. (Carlstein 1992). The major drawbacks of such resampling techniques are in general the additional compuational complexity and the lack of strong theoretical results on their statistical properties.
- For the case without regularization and constraints, there are numerous closely related techniques such as the FPE (Final Prediction Error) statistic (Akaike 1969), the AIC (Akaike Information Criterion) (Akaike 1974), the MDL (Minimum Description Length) statistic (Rissanen 1978), the GCV (Generalized Cross-Validation) statistic (Craven and Wahba 1979) and other approximative cross-validation statistics (Stoica *et al.* 1986). The all make use of the average residuals, but make a correction for the dependence between the residuals using asymptotic considerations.

The main purpose of this paper is to discuss the extension of the classical FPE statistic

$$\mathrm{FPE}(Z^N) = \frac{1 + d/N}{1 - d/N} V_N(\hat{\theta}_{PE}; Z^N)$$

derived from the standard prediction error method formulation to the generalized formulation with regularization and constraints.

[2] This formulation can be easily extended to the case with multiple penalties and regularization parameters

3.1 *FPE with Constraints*

First, we observe that the introduction of constaints (but no regularization) leads to a trivial modification of the FPE:

$$\mathrm{FPEC}(Z^N) = \frac{1+d_0/N}{1-d_0/N} V_N(\hat{\theta}_{PE}; Z^N)$$

where the degrees of freedom is $d_0 = d - d_a$. Here d_a is the number of linearly independent and active linearized constaints at the point $\hat{\theta}_{PE}$. This modification can be made rigorous by the application of the Implicit Function Theorem in order to reparameterize the criterion function at the point $\hat{\theta}_{PE}$ in terms of a d_0 dimensional parameter vector, as suggested in (Johansen 1996*a*).

3.2 *FPE with Regularization*

Second, an extension of the FPE to the case when there is regularization (but no constraints) can be found in (Larsen and Hansen 1994):

$$\mathrm{FPER}(\gamma; Z^N) = \frac{1+d_2(\gamma)/N}{1-2d_2(\gamma)/N + d_1(\gamma)/N} \cdot V_N(\hat{\theta}_{\mathrm{REG},\gamma}; Z^N)$$

where the two different expressions for the model's degrees of freedom are given by $d_1(\gamma) = \mathrm{tr}(S(\gamma))$ and $d_2(\gamma) = \mathrm{tr}(S(\gamma)S(\gamma))$, where

$$\begin{aligned} S(\gamma) &= \left(H_N(\hat{\theta}_{\mathrm{REG},\gamma}) + \gamma H_\Omega(\hat{\theta}_{\mathrm{REG},\gamma})\right)^{-1} \\ &\quad \cdot H_N(\hat{\theta}_{\mathrm{REG},\gamma}) \\ H_N(\theta) &= \nabla_\theta^2 V_N(\theta; Z^N) \\ H_\Omega(\theta) &= \nabla_\theta^2 \Omega(\theta) \end{aligned}$$

3.3 *FPE with Constraints and Regularization*

Finally, we discuss the case when both regularization and constraints are being applied. The idea is to reparameterize the FPER statistic in terms of a lower-dimensional parameter vector by eliminating paramteres corresponding to the active constraints, and compute the value of the reparameterized FPER criterion. This was suggested in (Johansen 1996*a*), and the details follow.

In a neighborhood of $\hat{\theta}_{\mathrm{REG},\gamma}$ the constraints are approximation by their linearizations:

$$A_G\theta - b_G = 0, \quad A_K\theta - b_K \leq 0$$

where

$$\begin{aligned} A_G &= \nabla_\theta G(\hat{\theta}_{\mathrm{REG},\gamma}) \\ A_K &= \nabla_\theta K(\hat{\theta}_{\mathrm{REG},\gamma}) \\ b_G &= \nabla_\theta G(\hat{\theta}_{\mathrm{REG},\gamma})\hat{\theta}_{\mathrm{REG},\gamma} - G(\hat{\theta}_{\mathrm{REG},\gamma}) \\ b_K &= \nabla_\theta K(\hat{\theta}_{\mathrm{REG},\gamma})\hat{\theta}_{\mathrm{REG},\gamma} - K(\hat{\theta}_{\mathrm{REG},\gamma}) \end{aligned}$$

Keeping only the active and linearly independent constraints at $\hat{\theta}_{\mathrm{REG},\gamma}$, it is clear that $\hat{\theta}_{\mathrm{REG},\gamma}$ satisfies:

$$A\hat{\theta}_{\mathrm{REG},\gamma} - b = 0$$

where $\dim(A) = d_a \times d_a$, and A contains the active and linearly independent rows of A_G and A_K. Likewise, b contains the corresponding elements of b_G and b_K. Without loss of generality[3], it can be assumed that the d-dimensional vector θ can be decomposed into a pair of d_a and $(d - d_a)$-dimensional sub-vectors θ_1 and θ_2 and the A matrix into two sub-matrices A_1 and A_2 such that

$$\begin{pmatrix} A_1 & A_2 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = b$$

$\dim(A_1) = d_a \times d_a$, $\dim(A_2) = d_a \times (d - d_a)$, and $\mathrm{rank}(A_1) = d_a$. Now, it is clear that θ_1 satisfies

$$\theta_1 = A_1^{-1}(b - A_2\theta_2)$$

Hence, an approximate (linearized) reparameterization of θ in terms of the $(d - d_a)$-dimensional vector θ_2 is defined by

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} A_1^{-1}(b - A_2\theta_2) \\ \theta_2 \end{pmatrix} = P\theta_2 - q$$

where the last inequality is an implicit definition of P and q. Notice that the above approximation is exact at the point $\hat{\theta}_{\mathrm{REG},\gamma}$. Defining

$$\hat{\theta}^T_{\mathrm{REG},\gamma} = \left(\hat{\theta}^T_{1,\mathrm{REG},\gamma}, \ \ \hat{\theta}^T_{2,\mathrm{REG},\gamma}\right)$$

it is clear that

$$\begin{aligned} V_N(\hat{\theta}_{\mathrm{REG},\gamma}; Z^N) &= V_N(P\hat{\theta}_{2,\mathrm{REG},\gamma} - q; Z^N) \\ &= \tilde{V}_N(\hat{\theta}_{2,\mathrm{REG},\gamma}) \end{aligned}$$

where the last inequality is a definition of the reparameterized criterion $\tilde{V}_N$. Now we are in position of apply the FPER statistic to define FPECR since the constraints are removed in the re-parameterized version of the criterion:

$$\mathrm{FPECR}(\gamma; Z^N) = \frac{1+\tilde{d}_2(\gamma)/N}{1-2\tilde{d}_2(\gamma)/N + \tilde{d}_1(\gamma)/N} \cdot V_N(\hat{\theta}_{\mathrm{REG},\gamma}; Z^N)$$

[3] The ordering of the elements in the vector θ can be rearranged to achieve this.

where the two different expressions for the model's degrees of freedom are given by $\tilde{d}_1(\gamma) = \mathrm{tr}(\tilde{S}(\gamma))$ and $\tilde{d}_2(\gamma) = \mathrm{tr}(\tilde{S}(\gamma)\tilde{S}(\gamma))$. Moreover,

$$\tilde{S}(\gamma) = \left(\tilde{H}_N(\hat{\theta}_{2,\mathrm{REG},\gamma}) + \gamma \tilde{H}_\Omega(\hat{\theta}_{2,\mathrm{REG},\gamma})\right)^{-1} \cdot \tilde{H}_N(\theta_{2,\mathrm{REG},\gamma})$$

where

$$\tilde{H}_N(\theta_2) = \nabla^2_{\theta_2} \tilde{V}_N(\theta_2; Z^N)$$
$$\tilde{H}_\Omega(\theta_2) = \nabla^2_{\theta_2} \tilde{\Omega}(\theta_2)$$

3.4 *Discussion*

The derivation of the classical Final Prediction Error is based on first order Taylor-expansions (linearization) (Söderström and Stoica 1988). The error introduced by the linearization of the contraints are therefore expected to be of the same order as the other approximation error in the derivations. The asymptotic statistical properties of the FPECR criterion is therefore expected to be similar to those of the FPER criterion, see (Larsen and Hansen 1994).

The above criteria can be easily generalized to a MIMO system identification framework, including multiple regularization penalties with separate regularization parameters.

Some of the above methods for regularization and FPE-like statistics are implemented as part of a computer-aided modeling tool for developing operating regime based models (Johansen 1996*b*, Johansen and Foss 1997), which are closely related to Takagi-Sugeno-Kang fuzzy models (Takagi and Sugeno 1985).

4. CONCLUDING REMARKS

A general framework for introducing prior knowledge in terms of penalties (including regularization) and constraints into non-linear system identification problems was described in (Johansen 1996*a*). In the present paper, the FPE statistic is extended to handle the case when the prediction error criterion is augmented with penalties and constraints, leading to the FPECR. This statistic is useful both for model structure identification and selection of regularization parameters.

5. REFERENCES

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Stat. Math.* **21**, 243–247.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control* **19**, 716–723.

Bai, E. W. and S. S. Sastry (1986). Parameter identification using prior information. *Int. J. Control* **44**, 455–473.

Carlstein, E. (1992). Resampling techniques for stationary time-series: Some recent developments. In: *New Direction in Time Series Analysis. Part I* (D. Brillinger et al., Ed.). pp. 75–85. Springer-Verlag, New York, NY.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Math.* **31**, 317–403.

Dayal, B. S. and J. F. MacGregor (1996). Identification of finite impulse response models: Methods and robustness issues. *Industrial and Engineering Chemistry Research* **35**, 4078–4090.

De Moor, B., M. Gevers and G. C. Goodwin (1994). l_2-overbiased, l_2-underbiased and l_2-unbiased estimation of transfer functions. *Automatica* **30**, 893–898.

Eskinat, E. (1995). System identification using constrained estimation. In: *Proc. European Control Conference, Rome.* pp. 856–861.

Eskinat, E., S. Johnson and W. Luyben (1993). Use of auxiliary information in system identification. *Ind. Eng. Chem. Research* **32**, 1981–1992.

Foss, B. A. and T. A. Johansen (1997). Identification and convexity in optimizing control. To be presented at the IFAC Symposium on System Identification, Kitakyushu, Japan.

Gawthrop, P. J., R. W. Jones and S. A. MacKenzie (1992). Identification of partially-known systems. *Automatica* **28**, 831–836.

Johansen, T. A. (1996*a*). Identification of non-linear systems using empirical data and prior knowledge – An optimization approach. *Automatica* **32**, 337–356.

Johansen, T. A. (1996*b*). Robust identification of Takagi-Sugeno-Kang fuzzy models using regularization. In: *Proc. IEEE Conf. Fuzzy Systems, New Orleans.* pp. 180–186.

Johansen, T. A. (1997). On Tikhonov regularization, bias and variance in nonlinear system identification. *Automatica* **33**, 441–446.

Johansen, T. A. and B. A. Foss (1997). ORBIT - operating regime based modeling and identification toolkit. To be presented at the IFAC Symposium on System Identification, Kitakyushu, Japan.

Kárný, M., A. Halousková and P. Nedoma (1995). Recursive approximation by ARX model: A tool for grey-box modelling. *Int. J. Adaptive Control and Signal Processing* **9**, 525–546.

Kramer, M. A., M. L. Thompson and P. M. Phagat (1992). Embedding theoretical models in neural networks. In: *Proceedings American Control Conference, Chicago, Il.*. pp. 475–479.

Kunisch, K. and E. W. Sachs (1992). Reduced SQP methods for parameter identification problems. *SIAM J. Numerical Analysis* **29**, 1793–1820.

Larsen, J. and L. K. Hansen (1994). Generalization performance of regularized neural network models. In: *Proc. IEEE Workshop on Neural Networks for Signal Processing, Ermioni, Greece.*

Moons, C. and B. De Moor (1995). Parameter identification of induction motor drives. *Automatica* **31**, 1137–1147.

Peterka, V. (1981). Bayesian system identification. *Automatica* **17**, 41–53.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **14**, 465–471.

Sjöberg, J., H. Hjalmarsson and L. Ljung (1994). Neural networks in system identification. In: *Preprints 10th IFAC Symp. System Identification, Copenhagen.* Vol. 2. pp. 49–72.

Sjöberg, J., T. McKelvey and L. Ljung (1993). On the use of regularization in system identification. In: *Preprints 12th IFAC World Congress, Sydney.* Vol. 7. pp. 381–386.

Söderström, T. and P. Stoica (1988). *System Identification.* Prentice Hall, Englewood Cliffs, NJ.

Stoica, P., P. Eykhoff, P. Janssen and T. Söderström (1986). Model-structure selection by cross-validation. *Int. J. Control* **43**, 1841–1878.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Royal Statistical Soc. B* **36**, 111–133.

Takagi, T. and M. Sugeno (1985). Fuzzy identification of systems and its application to modeling and control. *IEEE Trans. Systems, Man, and Cybernetics* **15**, 116–132.

Thompson, M. L. and M. A. Kramer (1994). Modeling chemical processes using prior knowledge and neural networks. *AIChE J.* **40**, 1328–1340.

Tikhonov, A. N. and V. Y. Arsenin (1977). *Solutions of Ill-posed Problems.* Winston, Washington DC.

Tulleken, H. J. A. F. (1993). Grey-box modelling and identification using physical knowledge and Bayesian techniques. *Automatica* **29**, 285–308.

Xin, J., H. Ohmori and A. Sano (1995). Minimum MSE based regularization for system identification in the presence of input and output noise. In: *Proc. 34th IEEE Conf. Decision and Control, New Orleans.* pp. 1807–1814.