On Tikhonov Regularization, Bias and Variance in Nonlinear System Identification

Tor A. Johansen

SINTEF Electronics and Cybernetics, Automatic Control Department, N-7034 Trondheim, Norway. Email: Tor.Arne.Johansen@ecy.sintef.no. Fax: +47 73 59 43 99. Phone: +47 73 59 03 95.

Regularization is a general method for solving ill-posed and ill-conditioned problems. Traditionally, ill-conditioning in system identification problems is usually approached using regularization methods such as ridge regression and principal component regression. In this work it is argued that the Tikhonov regularization method is a powerful alternative for regularization of non-linear system identification problems by introducing smoothness of the model as a prior. Its properties is discussed in terms of an analysis of bias and variance, and illustrated by a semi-realistic simulation example.

Key words: Regularization; System Identification; Nonlinear Systems; Asymptotic Approximation; Bias-Variance Trade-off.

To appear in *Automatica.*

1 Introduction

When the model structure does not match the system, is poorly identifiable, or the available set of empirical data is not sufficiently informative, the parameter identification problem may be ill-conditioned. If a standard prediction error method is applied to fit the parameters, such a situation is characterized by a low sensitivity of the identification criterion in a sub-manifold of the parameter space. The excessive degrees of freedom corresponding to this sub-manifold can therefore be chosen more or less arbitrarily, and the estimate may be dominated by noise, collinearities or other data deficiencies. This may lead to a model with poor extrapolation properties. In other words, the prediction error method is not robust in such cases. This is clearly undesirable, and there exists at least two general approaches to resolve the problem:

(i) The development of an alternative model structure with less degrees of freedom and a more suitable parameterization that matches the system better.

(ii) To regularize the identification algorithm by introducing constraints or penalties in order to attract the excessive degrees of freedom towards reasonable values (Tikhonov and Arsenin 1977), see also (Johansen 1996*b*).

The first case corresponds to explicitly reducing the number of parameters in the model structure. This is certainly the most common approach. In the second case, the parameterization remains the same, but the degrees of freedom (also called effective number of parameters) will be reduced due to soft or hard constraints on the parameter space that are more or less implicitly introduced by the regularization. In general, this results in a better conditioned problem, a more robust identification algorithm, and eventually a more accurate model. The potential benefit with regularization is that the model can be improved without modifying the model structure, which can be a laborious task. A potential drawback is that the computational complexity may be significantly increased.

In this paper we study Tikhonov regularization (Tikhonov and Arsenin 1977). While Tikhonov regularization has had significant impact on several branches of science and engineering dealing with ill-posed and inverse problems, in particular modeling and analysis of high-dimensional or distributed signals and data (Tikhonov and Arsenin 1977, O'Sullivan 1986, Wahba 1990, Poggio *et al.* 1985), there has been few applications to system parameter identification, with the exceptions of some work on distributed parameter identification (Kravaris and Seinfeld 1985). Regularization methods that are commonly applied in system identification include ridge regression and principal component regression (PCR) (Sjöberg and Ljung 1992, Sjöberg *et al.* 1993). In this paper it is jusitifed that significant improvements in the robustness and accuracy of parameter identification methods can be achieved with simple means using regularization, and justify that Tikhonov regularization may be a useful alternative to the more commonly applied methods.

2 Tikhonov Regularization

Suppose a model structure, i.e. a set of equations parameterized by a p-dimensional parameter vector θ, is given. The parameter vector can be estimated using a standard prediction error estimator, e.g. (Ljung 1987)

$$\hat{\theta}_{PE}(Z^N) = \arg\min_{\theta \in D_c} V_N(\theta; Z^N)$$
$$V_N(\theta; Z^N) = \frac{1}{N} \sum_{t=1}^{N} \varepsilon^2(t; \theta)$$

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 using 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, the data are not sufficiently informative, or the model structure is over-parameterized or fundamentally wrong, the prediction error method may not be robust. This problem was briefly discussed in the introduction, and will be addressed in more rigorously in Section 3. Regularization is a general method for improving the robustness of mathematical algorithms by imposing additional regularity constraints on the solution. A general approach for finding regularized solutions (Tikhonov and Arsenin 1977) is to introduce a penalty term in the prediction error criterion

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

where Ω is a stabilizer for the problem, and $\gamma > 0$ is a scalar regularization parameter. The idea is that the penalty term will attract any excessive degrees of freedom in the model structure towards reasonable regions of the parameter space. Excessive degrees of freedom are characterized by a low sensitivity of the basic prediction error criterion $V_N(\theta; Z^N)$ with respect to perturbations in the

[1] Formulating the method and analysis for with multi-step-ahead predictors and systems with multiple inputs or outputs is straightforward.

corresponding sub-manifold of the parameter space. Hence, the penalty $\gamma\Omega(\theta)$ should contribute significantly to the criterion when θ is in this sub-manifold. The parameter estimate is now

$$\hat{\theta}_\gamma^{\mathrm{REG}}(Z^N) = \arg\min_{\theta \in D_c} \left(V_N(\theta; Z^N) + \gamma\Omega(\theta) \right).$$

The most common choice of penalty is the ridge regression stabilizer (Levenberg 1944, Marquardt 1963, Hoerl and Kennard 1970, Swindel 1976) $\Omega_{\mathrm{LM}}(\theta) = (\theta - \theta^{\#})^T(\theta - \theta^{\#})$ which attracts the parameter estimate towards a point $\theta^{\#}$ in the parameter space. This is a simple, but widely applied and powerful technique.

A description of the Tikhonov regularization method follows. Assume the predictor is "parameterized" by a function f that we assume is smooth and parameterized by θ. For example, consider the state-space model

$$\begin{aligned} x(t+1) &= f(x(t), u(t), \theta) + v(t) \\ y(t) &= h(x(t)) + w(t) \end{aligned}$$

where $x(t)$ is the system state, $v(t)$ and $w(t)$ are disturbances and noise, and f and h are functions. On the basis of this model, a predictor $\hat{y}(t|Z^{t-1}, \theta)$ can be formulated, using for instance an extended Kalman-filter, e.g. (Ljung 1987). A possible formulation of the Tikhonov stabilizer is now

$$\Omega'_{\mathrm{TIK}}(\theta) = \int_\xi \left|\left| \nabla_\xi^2 f(\xi, \theta) \right|\right|_F^2 \rho(\xi) d\xi \tag{1}$$

where $\xi = (x, u)$ and the norm $||\cdot||_F$ is the Frobenius matrix norm defined by $||A||_F = \sqrt{\sum_{ij} a_{ij}^2}$. The positive semi-definite function ρ must reflect the operating region of the system, the intended application of the model and a priori knowledge about smoothness. A more general Tikhonov-stabilizer can contain weighted mixtures of derivatives of different orders. This approach is generally applicable for a wide range of model representations other than state-space models. For example, Kravaris and Seinfeld (1985) use a similar approach for identification of distributed parameter models. Moreover, it is clear how to extend the stabilizer to the case when the model contains multiple functions that parameterizes the predictor.

The Tikhonov stabilizer has at least two appealing interpretations that are relevant for the parameter identification problem. First, it favors models that give predictions that are as smooth functions of the past data as possible, since the norm of the Hessian (curvature in the scalar case) can be used as an intuitive measure of smoothness of a function. This is an attractive property, because smooth behavior is a reasonable prior or desired property of the model in many modeling and identification problems. Second, it attracts the parameters towards the sub-manifold or sub-set of the parameter space that corresponds to all linear predictors, since the Hessian of the predictor is zero everywhere for a linear predictor. Hence, when there are excessive degrees of freedom, the predictor will be attracted locally towards the linear one that is most consistent with the observations, locally. Again, this is reasonable, because many systems can be accurately described with linear predictors locally. Of course, this is related to the smoothness properties of the system.

3 Bias, Variance, and Prior Knowledge

In this section the effect of regularization on the identified model is investigated, both in terms of the accuracy of the parameter estimate (which may be of particular importance with physical models), but also in terms of the mean squared prediction error (which is of major importance with black box models). In particular, we study the asymptotic bias and variance when regularization is applied, and compare to the prediction error method.

First, some notation must be introduced. Assume the output is made up of a predictable term $g(Z^{t-1})$ and an unpredictable random term $e(t)$ with zero mean and finite variance

$$y(t) = g(Z^{t-1}) + e(t).$$

Let E denote expectation with respect to the joint probability distribution of all stochastic processes. Like in (Ljung 1987) we define the operator $\overline{E}s(t) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} Es(t)$ where s is a stochastic process, and it is implicitly assumed that the limit exists in all subsequent applications of this operator. Like Ljung (1987) we also define $\overline{V}(\theta) = \overline{E}\varepsilon^2(t;\theta)$, and assume that there exists a unique optimal parameter vector $\theta^\star$ that minimizes this function, i.e. $\theta^\star = \arg\min_{\theta\in D_c} \overline{V}(\theta)$. Finally, suppose the system and models under study satisfies certain growth constraints, and are exponentially stable, i.e. the remote past is forgotten at an exponential rate, in the sense of (Ljung 1978).

3.1 Parameter estimator accuracy

Under the conditions outlined above, it is well known that $\hat{\theta}_{PE}(Z^N) \to \theta^\star$ with probability one as $N \to \infty$, see (Ljung 1978). Hence, the prediction error method is asymptotically unbiased. Moreover, it is also well known that the estimate is asymptotically normally distributed

$$\sqrt{N}\left(\hat{\theta}_{PE}(Z^N) - \theta^\star\right) \stackrel{\text{distr}}{\longrightarrow} \mathcal{N}(0, P_\theta) \quad \text{as } N \to \infty$$

under similar general conditions (Ljung and Caines 1979), where the covariance matrix is

$$P_\theta = \left(\overline{H}(\theta^\star)\right)^{-1} Q \left(\overline{H}(\theta^\star)\right)^{-1}$$

where $\overline{H}(\theta) = \nabla_\theta^2 \overline{V}(\theta)$ and $Q = \lim_{N\to\infty} E\left(N \nabla_\theta V_N(\theta^\star; Z^N) \nabla_\theta V_N^T(\theta^\star; Z^N)\right)$. It is assumed that the Hessian $\overline{H}(\theta^\star)$ is positive definite. The asymptotic expression for the covariance matrix equals the asymptotic limit of the Cramer-Rao lower bound, provided the residuals are assumed to be Gaussian, e.g. (Ljung and Caines 1979). It is clear that the prediction error estimate $\hat{\theta}_{PE}(Z^N)$ is asymptotically the best unbiased estimate of $\theta^\star$. However, that does not necessarily mean that it is very accurate or even useful. For instance, there exists parameter identification problems where the Hessian $\overline{H}(\theta^\star)$ is ill-conditioned. A direct consequence of the ill-conditioning of $\overline{H}(\theta^\star)$ is large variance in some sub-space of the parameter space.

Let us return to the regularized version of the prediction error method. For a general stabilizer Ω, we define

$$\overline{V}_\gamma^{\text{REG}}(\theta) = \overline{V}(\theta) + \gamma\Omega(\theta)$$

Hence, the minima of the regularized and non-regularized asymptotic prediction error criteria ($\theta_\gamma^\star$ and $\theta^\star$ respectively) satisfy $\nabla_\theta \overline{V}(\theta_\gamma^\star) + \gamma \nabla_\theta \Omega(\theta_\gamma^\star) = 0$ and $\nabla_\theta \overline{V}(\theta^\star) = 0$, respectively. Taylor-expanding the gradient $\nabla_\theta \overline{V}_\gamma^{\text{REG}}$ gives

$$\nabla_\theta \overline{V}_\gamma^{\text{REG}}(\theta^\star) = \nabla_\theta \overline{V}_\gamma^{\text{REG}}(\theta_\gamma^\star) + \nabla_\theta^2 \overline{V}_\gamma^{\text{REG}}(\theta_\gamma^\star)(\theta^\star - \theta_\gamma^\star) \;+\; \text{h.o.t.}$$

and neglecting the higher order terms, we get

$$\theta_\gamma^\star = \theta^\star - \gamma\left(\overline{H}(\theta_\gamma^\star) + \gamma H_\Omega(\theta_\gamma^\star)\right)^{-1} \nabla_\theta \Omega(\theta^\star)$$

where we have assumed that the Hessian $H_\Omega(\theta_\gamma^\star) = \nabla_\theta^2 \Omega(\theta_\gamma^\star)$ is positive definite, but in contrast to above, we do not require that $\overline{H}(\theta_\gamma^\star)$ is positive definite. Hence, under the previously stated assumptions, it follows from

$$\hat{\theta}_\gamma^{\mathrm{REG}}(Z^N) - \theta^\star = \left(\hat{\theta}_\gamma^{\mathrm{REG}} - \theta_\gamma^\star\right) + \left(\theta_\gamma^\star - \theta^\star\right)$$

and the result of (Ljung and Caines 1979) that

$$\sqrt{N}\left(\hat{\theta}_\gamma^{\mathrm{REG}}(Z^N) - \theta^\star\right) \stackrel{\mathrm{distr}}{\longrightarrow} \mathcal{N}(b_\gamma, P_{\theta,\gamma}^{\mathrm{REG}}) \quad \text{as } N \to \infty$$

where the asymptotic expressions for the bias and covariance matrix are

$$b_\gamma = -\gamma\left(\overline{H}(\theta_\gamma^\star) + \gamma H_\Omega(\theta_\gamma^\star)\right)^{-1} \nabla_\theta \Omega(\theta^\star) \tag{2}$$

$$P_{\theta,\gamma}^{\mathrm{REG}} = \left(\overline{H}(\theta_\gamma^\star) + \gamma H_\Omega(\theta_\gamma^\star)\right)^{-1} Q \left(\overline{H}(\theta_\gamma^\star) + \gamma H_\Omega(\theta_\gamma^\star)\right)^{-1} \tag{3}$$

respectively. We observe from (2) and (3) that for $\gamma > 0$, the regularized estimator may be biased, and the covariance matrix of the regularized estimator will be less than the Cramer-Rao lower bound, since $\gamma H_\Omega(\theta_\gamma^\star)$ is positive definite. Since the estimator may be biased, we are more interested in the total error

$$\tilde{P}_{\theta,\gamma}^{\mathrm{REG}} = E\left(\hat{\theta}_\gamma^{\mathrm{REG}}(Z^N) - \theta^\star\right)\left(\hat{\theta}_\gamma^{\mathrm{REG}}(Z^N) - \theta^\star\right)^T$$

rather than the covariance matrix $P_{\theta,\gamma}^{\mathrm{REG}}$. It is evident that the regularized estimator will be more accurate than the unregularized prediction error estimator provided the decrease in accuracy due to the bias is less than the increase in accuracy due to the reduced variance. Hence, the biased regularized estimator may be more accurate than the best unbiased one; we recognize the well-known trade-off between bias and variance.

3.2 The mean squared prediction error

The mean squared prediction error is defined by

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

for some arbitrary estimator $\hat{\theta}(Z^N)$. The variables Z^N, Z^{t-1}, and $e(t)$ are viewed as stochastic, so MSE is the ensemble average over all possible identification data sequences of length N, and future data sequences. Since the stochastic process e is white noise with variance $\overline{E}e^2(t) = \lambda_0$, and the future data are independent of the identification data, we get

$$\begin{aligned}\text{MSE} = \overline{E}\left(g(Z^{t-1}) - E\left(\hat{y}(t;Z^{t-1},\hat{\theta}(Z^N))\middle| Z^{t-1}\right)\right)^2 \\ +\overline{E}\left(\hat{y}(t;Z^{t-1},\hat{\theta}(Z^N)) - E\left(\hat{y}(t;Z^{t-1},\hat{\theta}(Z^N))\middle| Z^{t-1}\right)\right)^2 + \lambda_0 \end{aligned} \quad (4)$$

This is a bias/variance decomposition of the expected squared prediction error. The first term is the systematic error (squared bias). The second term is the random error (variance), which arises because the optimal parameters cannot in general be exactly determined on the basis of the finite data sequence Z^N. The third term is the variance of the noise.

Recall that for the prediction error method without any regularization, we have the following asymptotic expression for the variance part of the MSE

$$\overline{E}\left(\hat{y}(t;Z^{t-1},\hat{\theta}_{PE}(Z^N)) - E\left(\hat{y}(t;Z^{t-1},\hat{\theta}_{PE}(Z^N))\middle| Z^{t-1}\right)\right)^2 = \frac{p}{N}\lambda_0 \quad (5)$$

where p is the number of parameters (Söderström and Stoica 1988). Eqs. (4) and (5) give the well known asymptotic expression

$$\text{MSE} = b^2 + \left(1 + \frac{p}{N}\right)\lambda_0 \quad (6)$$

where the squared bias is

$$b^2 = \overline{E}\left(g(Z^{t-1}) - E\left(\hat{y}(t;Z^{t-1},\hat{\theta}_{PE}(Z^N))\middle| Z^{t-1}\right)\right)^2$$

With a regularized prediction error method, we get the asymptotic expression (Larsen and Hansen 1994) (see also (Xin *et al.* 1995))

$$\overline{E}\left(\hat{y}(t;Z^{t-1},\hat{\theta}_\gamma^{\text{REG}}(Z^N)) - E\left(\hat{y}(t;Z^{t-1},\hat{\theta}_\gamma^{\text{REG}}(Z^N))\middle| Z^{t-1}\right)\right)^2 = \frac{d(\gamma)}{N}\lambda_0$$

and

$$\text{MSE}_\gamma^{\text{REG}} = b^2(\gamma) + \left(1 + \frac{d(\gamma)}{N}\right)\lambda_0 \quad (7)$$

where

$$b^2(\gamma) = \overline{E}\left(g(Z^{t-1}) - E\left(\hat{y}(t;Z^{t-1},\hat{\theta}_\gamma^{\text{REG}}(Z^N))\middle| Z^{t-1}\right)\right)^2$$

is the squared bias, and the degrees of freedom in the model is given by

$$d(\gamma) = \mathrm{tr}\left(\left(\overline{H}(\theta_\gamma^\star) + \gamma H_\Omega(\theta_\gamma^\star) \right)^{-1} Q \left(\overline{H}(\theta_\gamma^\star) + \gamma H_\Omega(\theta_\gamma^\star) \right)^{-1} Q \right).$$

It follows that $0 \leq d(\gamma) \leq p$, $d(0) = p$, and $d(\gamma) \to 0$ as $\gamma \to \infty$. Hence, if the length N of the data sequence Z^N is kept fixed, and the regularization parameter γ decreases, the bias will typically decrease, while the variance will increase, see Fig. 1. Hence, we have a similar tradeoff between bias and variance as described in section 3.1, which suggests that there exists a γ that gives a model with optimal degrees of freedom. A major problem is to select γ such that it gives the optimal balance between bias and variance. Notice that the optimal values of γ minimizing $\mathrm{MSE}_\gamma^{\mathrm{REG}}$ and the parameter estimator accuracy $\tilde{P}_{\theta,\gamma}$, in some sense, may in general be different. We will return to the problem of selecting γ in section 3.4.

3.3 Prior Knowledge

From sections 3.1 and 3.2 we make the following general observation: Regularization will reduce the variance of the parameter estimate and MSE at the possible cost of a bias. The "magnitude" of the bias will depend on the prior knowledge incorporated in the stabilizer Ω whose purpose is to attract the excessive degrees of freedom towards reasonable regions of the parameter space. Hence, the more prior knowledge about the reasonable regions of the parameter space coded into Ω, the more we can reduce the variance without introducing significant bias. An attractive property of regularization is that it provides a general framework for incorporation of prior knowledge into the identification algorithm (Johansen 1996*b*). Another important observation is that when the parameter identification problem is ill-conditioned, then significant improvements can be achieved using very weak prior knowledge like $\theta^\# = 0$ with ridge regression, or an assumption about smoothness with Tikhonov regularization. The reason is that we can achieve great reduction of variance at the cost of a small bias. With the close relationship between regularization and the introduction of prior knowledge, it should be no surprise that regularized estimation can be framed as a Bayesian method, see e.g. (MacKay 1991).

3.4 Choosing the Regularization Parameter

The bias/variance trade-off interpretation illustrates how regularization can be applied to find the model with the best prediction performance. If this reflects the main purpose of the parameter identification, it is natural to choose the regularization parameter that gives the best trade-off between bias and variance, i.e. minimizes MSE. Methods for estimating the MSE include the

Final Prediction Error criterion for Regularized models (FPER) (Larsen and Hansen 1994)

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

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

$$S(\gamma) = \left(H_N(\hat{\theta}_\gamma^{\mathrm{REG}}) + \gamma H_\Omega(\hat{\theta}_\gamma^{\mathrm{REG}})\right)^{-1} H_N(\hat{\theta}_\gamma^{\mathrm{REG}})$$

and $H_N(\theta) = \nabla_\theta^2 V_N(\theta; Z^N)$. This criterion is a generalization of the well known Final Prediction Error (FPE) criterion, e.g. (Ljung 1987). We suggest to choose the regularization parameter γ that minimizes FPER:

$$\hat{\gamma}(Z^N) = \arg\min_{\gamma \geq 0} \quad \mathrm{FPER}(\gamma; Z^N)$$

At least a local minimum can be found using a simple line search algorithm. Extensions of the FPER statistic to also cover identification with both regularization and constraints can be found in (Johansen 1996*a*).

Alternative approaches includes the use of a separate validation data sequence, data re-sampling techniques like bootstrapping (Carlstein 1992) or cross-validation (Stoica *et al.* 1986) to estimate the MSE. It should be remembered that all these approaches are computationally expensive since a nonlinear programming problem must potentially be solved for each γ in the seach.

4 Simulation Example

This simulation example is based on the results reported in (Foss *et al.* 1995), where model based control of a batch fermentation reactor is studied. The simulated "true system" model describes the fermentation of glucose to gluconic acid by the micro-organism *Pseudomonas ovalis* in a well-stirred batch reactor:

$$\begin{aligned}
\dot{x}_1 &= \mu_m \frac{x_1 x_4 x_5}{K_s x_5 + K_0 x_4 + x_4 x_5} \\
\dot{x}_2 &= v_L \frac{x_1 x_4}{K_L + x_4} - 0.9082 K_p x_2 \\
\dot{x}_3 &= K_p x_2
\end{aligned}$$

$$\dot{x}_4 = -\frac{1}{Y_s}\mu_m \frac{x_1 x_4 x_5}{K_s x_5 + K_0 x_4 + x_4 x_5} - 1.011 v_L \frac{x_1 x_4}{K_L + x_4}$$
$$\dot{x}_5 = k_l a(x_5^* - x_5) - 0.09 v_L \frac{x_1 x_4}{K_L + x_4} - \frac{1}{Y_0}\mu_m \frac{x_1 x_4 x_5}{K_s x_5 + K_0 x_4 + x_4 x_5}$$

where x_1 is the cell concentration, x_2 is gluconolactone concentration, x_3 is gluconic acid concentration, x_4 is glucose concentration and x_5 is dissolved oxygen concentration. The parameters μ_m, K_L, v_L, and K_p depend on the controlled temperature and pH (denoted u_1 and u_2, respectively). This dependency is given by an interpolated lookup table based on the experimental data. Initial values for the batch are $x_1(0) = x_{10}, x_2(0) = 0, x_3(0) = 0$ and $x_4(0) = x_{40}$. The initial value $x_5(0)$ is determined by the other initial values.

The model structure developed in (Foss *et al.* 1995) is based on an operating regime decomposition of the full operating range of the process, and the use of simple local linear models within each operating regime. These local linear models are interpolated according to the operating point in order to get a complete global non-linear prediction model. All the local models are chosen to have the same linear structure

$$x(t+1) = a_i + A_i x(t) + B_i u(t) \tag{8}$$

where $x = (x_1, ..., x_5)$, $u = (u_1, u_2)$ and

$$a_i = \begin{pmatrix} a_1^i \\ a_2^i \\ a_3^i \\ a_4^i \\ a_5^i \end{pmatrix}, \quad A_i = \begin{pmatrix} A_{11}^i & 0 & 0 & A_{14}^i & A_{15}^i \\ A_{21}^i & A_{22}^i & 0 & A_{24}^i & 0 \\ 0 & A_{32}^i & 1 & 0 & 0 \\ A_{41}^i & 0 & 0 & A_{44}^i & A_{45}^i \\ A_{51}^i & 0 & 0 & A_{54}^i & A_{55}^i \end{pmatrix}, \quad B_i = \begin{pmatrix} B_{11}^i & B_{12}^i \\ B_{21}^i & B_{22}^i \\ B_{31}^i & B_{32}^i \\ B_{41}^i & B_{42}^i \\ B_{51}^i & B_{52}^i \end{pmatrix}$$

The structural zeros follow from a simple mass-balance based on the reaction mechanism and the assumption that the reaction rates only depend on x_4 and x_5, in addition to u_1 and u_2. This is a quite natural assumption to make, since these are the rate-limiting components. By qualitatively examining the main reaction mechanisms, 16 operating regimes are selected, see (Foss *et al.* 1995) for details. This leads to 448 unknown model parameters that must be identified. With the above mentioned model structure, the one-step-ahead predictor is linearly parameterized. The prediction error method is therefore equivalent to the least squares algorithm, and the regularized algorithms also have very simple implementations since the criteria are quadratic.

In (Foss *et al.* 1995) the 448 parameters were found using the standard least-squares method, and simulated data from 600 batches, each run for 10 h, and all states "measured" every 0.5 h. For every batch, the initial states x_{10} and x_{40} were randomly chosen. Small disturbances on the stirring dynamics (the $k_l a$ term) and random measurement noise were added to the simulated data. The control input trajectories were designed by randomly selecting between 0 and 2 step changes, within the allowable ranges of both temperature and pH, during the batch. This parameter identification problem becomes ill-conditioned when the number of observations decreases. One reason for this is that the model has approximately the same complexity over a wide range of operating conditions, while the experiment design is not by any means optimal. This means that there will exist operating conditions from which there is very little data, while the model structure contains parameters that are exclusively related to these operating conditions.

The example compares the results of the least squares algorithm with various regularized versions, and in particular study the quality of the identified models as the number of data points decreases towards zero, which leads to an ill-conditioned and eventually ill-posed parameter identification problem. We have identified four models, using four different identification criteria. In all cases with regularization, the FPER criterion was used to select the regularization parameter $\gamma = \hat{\gamma}(Z^N)$.

(i) **Standard least squares criterion without regularization.**

(ii) **Least squares criterion with Tikhonov regularization.** Inserting the linearly parameterized predictor $\hat{y}(t; Z^{t-1}, \theta) = f(x(t-1), u(t-1); \theta) = \varphi^T(x(t-1), u(t-1))\theta$ in (1), the Tikhonov stabilizer can be written in the quadratic form $\Omega_{TIK}(\theta) = \theta^T S \theta$ with

$$S = \sum_{i=1}^{7} \sum_{j=1}^{7} \int_{\xi} \left(\frac{\partial^2}{\partial \xi_i \partial \xi_j} \varphi(\xi) \varphi^T(\xi) \right) \rho(\xi) d\xi$$

where $\rho(\xi)$ is 1 over the relevant operating range, and zero elsewhere. Hence, the minimization problem remains quadratic. In the implementation the derivatives and integral are computed using finite-difference numerical approximations. The uniform grid contains about 280.000 points.

(iii) **Least squares solved with ridge regression.** In particular, $\theta^{\#} = 0$.

(iv) **Least squares solved with PCR**. Singular value decomposition is used to invert the Hessian matrix in the least squares problem. Singular values less than a threshold are zeroed. This threshold is the regularization parameter.

The models will be compared by simulating a number of batches not used for identification. In these batches, the pH and temperature were randomly changed every 0.5 h. Moreover, the open loop optimal control performance with the identification models are also compared, using similar experiments as in (Foss *et al.* 1995). The average prediction error on the separate test data set is illustrated in Figure 2a as a function of the number of observations N in the identification data set Z^N. We clearly see that the standard least squares algorithm becomes ill-conditioned and finally ill-posed as N approaches the number of model parameters, i.e. 448, and eventually becomes smaller than 448. We also observe that the regularized versions of the algorithm degrade more gracefully and give useful models even when the number of observations is less than the number of parameters. Of course, with regularization it is the degrees of freedom rather than the number of parameters that is of interest, cf. Figure 2c. From this figure we see that the FPER algorithm found reasonable values of the regularization parameter giving a gradual reduction of the degrees of freedom in the model as the problem becomes more ill-conditioned. We also observe significant differences between the regularization methods, which are partially due to the different levels of bias introduced by the different regularization methods. We believe it is characteristic that Tikhonov regularization introduce less bias than for instance PCR, although we do not expect that this is always true. It should be remembered that Tikhonov regularization is equivalent to a smoothness prior, and may not work so well when the system is non-smooth and this is not reflected in the weighting function in the Tikhonov stabilizer. A different comparison of the identified models is provided in Figure 2b, which illustrates the control performance in terms of average production rate of gluconic acid with optimal controllers designed on the basis of the identified models. This also indicates the same ranking of the identified models.

5 Concluding Remarks

Lack of identifiability, or insufficient information in the available data sequence (lack of persistence of excitation) are typical reasons why a parameter identification problem may be ill-conditioned. In applications with highly complex, non-linear, and possibly over-parameterized models, ill-conditioning should be expected. The aim of regularization is to improve the robustness of the identification algorithm and the accuracy of the model by explicitly addressing the bias/variance trade-off and taking advantage of the fact that there may exist biased estimators that are better than the best unbiased one. Prior knowledge and bias are closely related: Many generic regularization penalties may give improvements, but application of prior knowledge in the stabilizer will give largest improvement because less bias is introduced. Our experience, including the simulation example, is that Tikhonov regularization is a powerful alternative for the identification of smooth non-linear models.

Acknowledgments

Thanks to Aage V. Sørsensen at The Norwegian University of Science and Technology for assistance with the simulation example.

References

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.

Foss, B. A., T. A. Johansen and Aa. V. Sørensen (1995). Nonlinear predictive control using local models - applied to a batch fermentation process. *Control Engineering Practice* **3**, 389–396.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* **12**, 55–62.

Johansen, T. A. (1996*a*). Constrained and regularized system identification. Preprint, submitted for publication.

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

Kravaris, C. and J. H. Seinfeld (1985). Identification of parameters in distributed parameter systems by regularization. *SIAM J. Control and Optimization* **23**, 217–241.

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.*

Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least squares. *Quart. Appl. Math.* **2**, 164–168.

Ljung, L. (1978). Convergence analysis of parametric identification methods. *IEEE Trans. Automatic Control* **23**, 770–783.

Ljung, L. (1987). *System Identification: Theory for the User.* Prentice-Hall, Inc., Englewood Cliffs, NJ.

Ljung, L. and P. E. Caines (1979). Asymptotic normality of prediction error estimation for approximate system models. *Stochastics* **3**, 29–46.

MacKay, D. J. C. (1991). Bayesian Methods for Adaptive Models. PhD thesis. California Institute of Technology, Pasadena, CA.

Marquardt, D. W. (1963). An algorithm for least squares estimation of nonlinear parameters. *J. SIAM* **11**, 431–441.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science* **1**, 502–527.

Poggio, T., V. Torr and C. Koch (1985). Computational vision and regularization. *Nature* **317**, 314–319.

Sjöberg, J. and L. Ljung (1992). Overtraining, regularization, and searching for minimum in neural networks. In: *Preprints IFAC Symposium on Adaptive Systems in Control and Signal Processing, Grenoble, France.* pp. 669–674.

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.

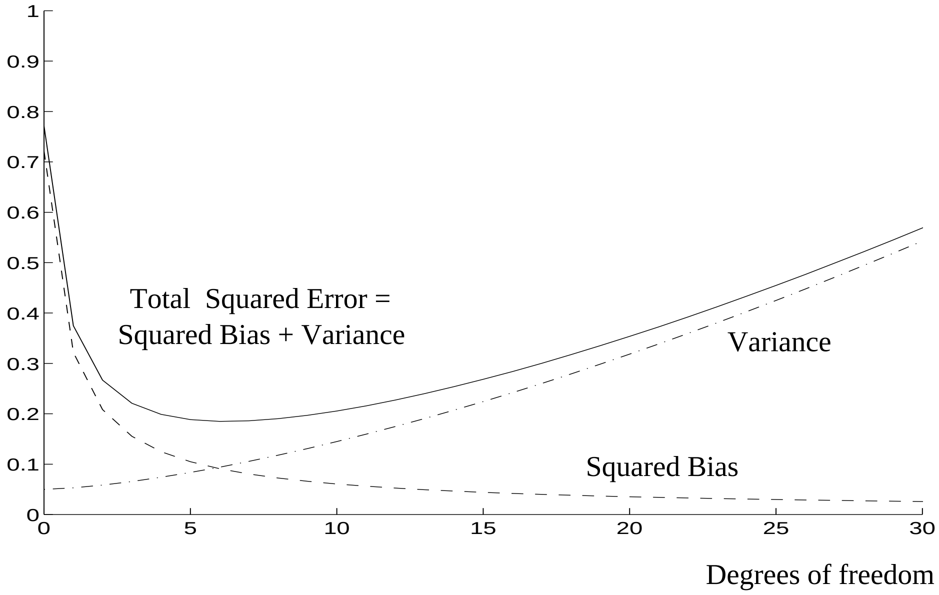
Stoica, P., P. Eykhoff, P. Janssen and T. Söderström (1986). Model-structure selection by cross-validation. *Int. J. Control* **43**, 1841–1878.

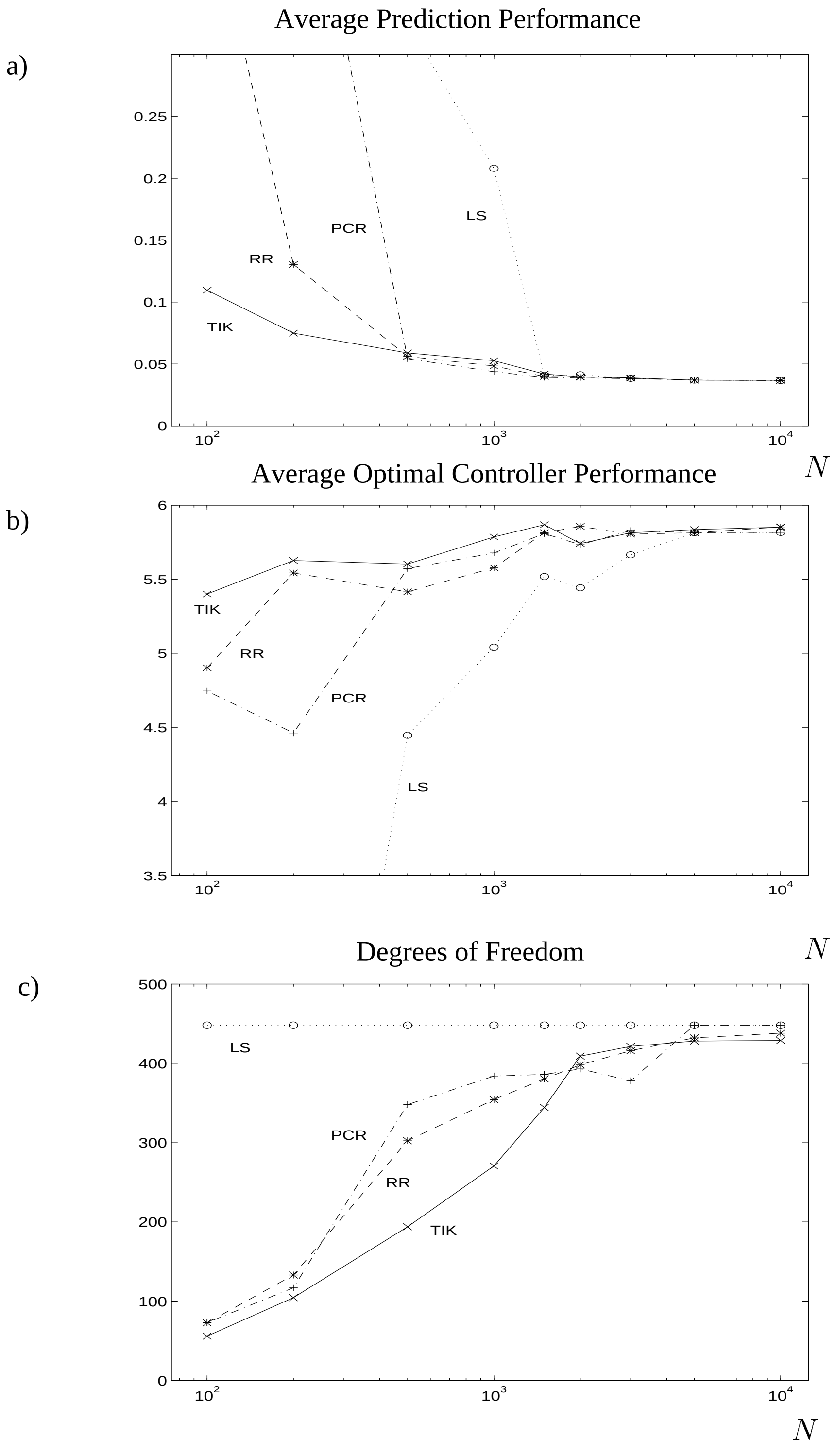
Swindel, B. F. (1976). Good ridge estimates based on prior information. *Comm. Statistics* **A5**, 985–997.

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

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

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.


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
30
Total Squared Error =
Squared Bias + Variance
Variance
Squared Bias
Degrees of freedom



Average Prediction Performance
a)
LS
PCR
RR
TIK
N
Average Optimal Controller Performance
b)
TIK
RR
PCR
LS
N
Degrees of Freedom
c)
LS
PCR
RR
TIK
N

Fig. 1. Typical relationship between total squared error, squared bias, variance and degrees of freedom in the model.

Fig. 2. Results of simulation experiments: a) Illustrates the estimated average prediction performance as a function of the number of observations N. b) Illustrates the optimal control performance in terms of average production rate of gluconic acid as a function of the number of observations N. c) Illustrates the degrees of freedom with the identified models as a function of the number of observations N. Models identified with the least squares algorithm are marked with o, models identified with PCR are marked with $+$, models identified using ridge regression are marked with $\star$, while models identified with Tikhonov regularization are marked with $\times$.