Robust Identification of Takagi-Sugeno-Kang Fuzzy Models using Regularization

Tor A. Johansen*

SINTEF Automatic Control,
N-7034 Trondheim, Norway.

Abstract

The identification of fuzzy models can sometimes be a difficult problem, often characterized by lack of data in some regions, collinearities and other data deficiencies, or a sub-optimal choice of model structure. Regularization is suggested as a general method for improving the robustness of standard parameter identification algorithms leading to more accurate and well-behaved fuzzy models. The properties of the method are related to the bias/variance tradeoff, and illustrated with a semi-realistic simulation example.

1 Introduction

The problem of identifying fuzzy models is often an ill-conditioned one, in the sense that there may be insufficient information in the empirical data to identify the model's parameters with sufficient accuracy. Fuzzy models, such as the Takagi-Sugeno-Kang model [24], are often characterized by a fuzzy partitioning of the system's operating range and a simple local model within each region. Poor accuracy or undesirable behavior of the identified model is often related to lack of a sufficient amount of data from some of the regions defined by the fuzzy partitioning. In such cases, it is obviously difficult to get good estimates and commonly used parameter identification methods like the least squares method are likely to fail for such identification problems. In other cases, the cause of the undesired model behavior may be less evident, as we shall see.

In this paper we will show that significant improvements in the robustness of fuzzy model identification algorithms can be achieved with simple means using regularization, and justify that the Tikhonov stabilizer may be a suitable stabilizer for fuzzy models. An approximate Tikhonov stabilizer that takes advantage of the particular structure of the TSK fuzzy model and is very fast to compute, is also presented.

Regularization has been applied with success with neural networks [1, 5] and computer vision [19]. It also plays a central role is statistics [18] and spline modeling [26]. It is also being applied in system identification [12, 20, 9].

*Email: Tor.Arne.Johansen@rektek.sintef.no

2 Least Squares Identification

Suppose the problem is to identify a function $f : X \rightarrow Y$ mapping inputs in $X \subset R^n$ to outputs in $Y \subset R$. The TSK fuzzy model consists of a set of rules

$$\begin{array}{l} \text{IF } x \text{ IS } X_1 \text{ THEN } y = f_1(x;\theta_1) \\ \quad\vdots \\ \text{IF } x \text{ IS } X_m \text{ THEN } y = f_m(x;\theta_m) \end{array}$$

where the fuzzy sets $X_1, ..., X_m$ define a complete fuzzy partition of X, and $f_1, ..., f_m$ are functions. Using the well known TSK fuzzy inference mechanism [24], the output of the fuzzy logic system is

$$f(x;\theta) = \sum_{i=1}^{m} f_i(x;\theta_i) w_i(x) \quad (1)$$

$$w_i(x) = \frac{\mu_{X_i}(x)}{\sum_{j=1}^{m} \mu_{X_j}(x)} \quad (2)$$

where μ_{X_i} is the membership function for the fuzzy set X_i. We assume here that the structure of the fuzzy model is given, and that the problem is one of identifying the local model parameters $\theta_1, ..., \theta_m$. We acknowledge that the structure identification problem is perhaps the more challenging and important one [22, 6, 27, 10], but as we shall see, there are some interesting problems with the parameter identification problem related to sub-optimal choice of model structure. For simplicity we also assume that the functions $f_1, ..., f_m$ are all equal to F, which is a linear functions of x

$$F(x;\theta_i) \quad = \quad \theta_{i,0} + \theta_{i,1}x_1 + ... + \theta_{i,n}x_n$$

We note that the general methods described here are applicable to any smooth local model structure F, but that for non-linearly parameterized structure F the criteria will not be quadratic and cannot always be minimized explicitly. On the basis of a set of empirical input/output data

$$Z^N = ((x(1), y(1)), (x(2), y(2)), ..., (x(N), y(N)))$$

we can now define the standard least squares identification criterion

$$\begin{aligned} V(\theta; Z^N) &= \frac{1}{N}\sum_{t=1}^{N}\left(y(t) - \sum_{i=1}^{m} F(x(t);\theta_i)w_i(x(t))\right)^2 \\ &= \frac{1}{N}(Y - \Phi\theta)^T(Y - \Phi\theta) \end{aligned}$$

which has the well known minimum

$$\hat{\theta}(Z^N) \quad = \quad (\Phi^T\Phi)^{-1}\Phi^T Y$$

and the definitions of Y, Φ and θ should be obvious. Next, we present some simple example of ill-conditioning leading to undesirable model behavior in the identification of fuzzy models using the least squares algorithm.

Consider first the case when there is very little data available from some of the regions defined by the fuzzy partitioning, see Figure 1. We see that the local models corresponding to these regions do not give as accurate description of the system as in the other regions with more data. Notice that in the extreme case when no data sample has input within the support of one of the membership functions, the problem is ill-posed and the parameters of the corresponding local model can be chosen arbitrarily, because they do not have any effect on the criterion.

While the above example is quite trivial, it should be noted that in practical problems with sparse data sets, high-dimensional systems, constraints on the experiment design and constraints on the relationships between the elements of the input vector x e.g. due to system dynamics or collinearities, this is a major problem. In particular, this will be the case if there is not sufficient a priori information about the system to select the structure of the fuzzy model to avoid these problem. Moreover, many popular structure identification algorithms [11] may give rise to fuzzy partitions

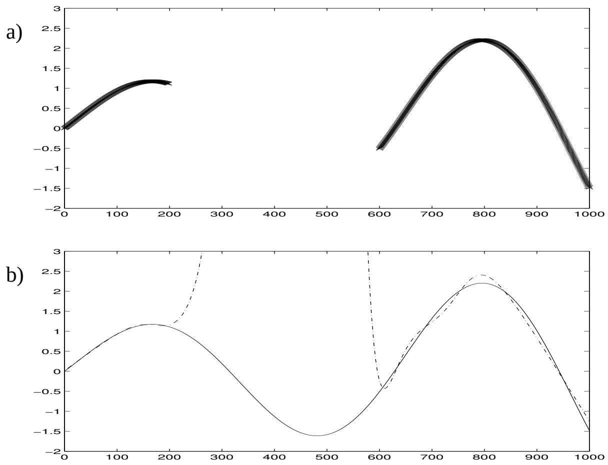

Figure 1. The 1000 noise-free observations in part a) are used to identify a TSK fuzzy model with five uniformly distributed local models. The result with least squares identification is shown in part b), dashed-dotted line, and compared to the true function.

without any guarantees on the amounts data in the resulting regions.

As a second example, consider a less obvious case where the distribution of data is approximately uniform over the operating range and the amount of data is very large compared to the number of parameters, see Figure 2 and [16, 17]. Also in this case we see that the local model parameters are characterized by high variability and the resulting model has undesirable behavior, namely the “oscillations”. As noted in [16, 17], the problem in this case is that the local model structure complexity is uniform (the fuzzy partitioning is uniform) but the system complexity is strongly non-uniform (due to the discontinuities in the function). This structural mismatch leads the identification algorithm to try to apply the local model in one region to describe phenomena in another region. This is possible because the Gaussian membership functions do not have compact support and therefore introduce interactions.

Other causes for ill-conditioning may be collinearities in the input data and other constraints on the input data. In dynamical system, lack of local persistence of excitation will also cause ill-conditioning for dynamic models of high order [21].

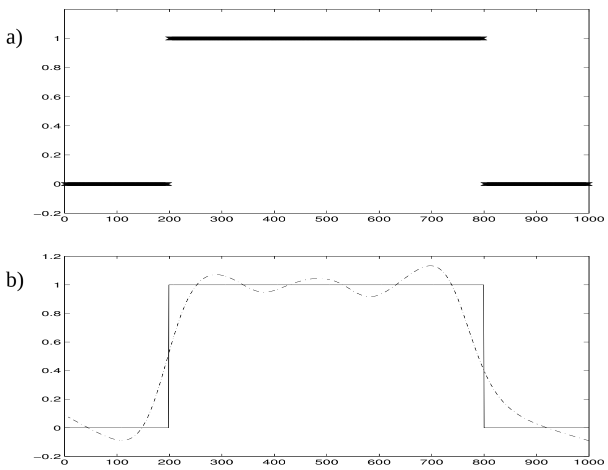

Figure 2. The 1000 noise-free observations in part a) are used to identify a TSK fuzzy model with five uniformly distributed local models. The result with least squares identification is shown in part b), dashed-dotted line, and compared to the true function.

3 The Method of Regularization

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 of the identification problem [25, 9] is to introduce a penalty term in the least squares criterion

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

where Ω is a stabilizer for the problem, and $\gamma > 0$ is a 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 criterion $V(\theta; Z^N)$ with respect to perturbations in the corresponding sub-space of the parameter space. Hence, the penalty $\gamma\Omega(\theta)$ should contribute significantly to the criterion when θ is in this sub-space.

3.1 Levenberg-Marquardt Regularization

The most common choice of penalty is the Levenberg-Marquardt stabilizer [14, 15]

$$\Omega_{\mathrm{LM}}(\theta) \quad = \quad (\theta - \theta^{\#})^T(\theta - \theta^{\#})$$

which means that the parameters are attracted towards a point $\theta^{\#}$ in the parameter space. This is a simple, but widely applied technique. Without significant prior knowledge, the choice $\theta^{\#} = 0$ is common. In statistics, it is often called ridge regression (see e.g. [7, 23]). Even though the optimal parameters may be significantly different than zero, this may lead to improvements in accuracy of the model compared to not using regularization at all.

3.2 Tikhonov Regularization

Let us proceed with a description of the Tikhonov regularization method [25]. A typical Tikhonov stabilizer is

$$\Omega_{\mathrm{TIK}}(\theta) \quad = \quad \int_{x \in X} \left|\left|\nabla_x^2 f(x;\theta)\right|\right|^2 dx \qquad (3)$$

where the norm $||\cdot||$ is the Frobenius matrix norm defined by $||A|| = \sqrt{\sum_{ij} a_{ij}^2}$. The more general Tikhonov-stabilizer can contain weighted mixtures of derivatives of different orders.

Inserting the linearly parameterized model into (3), the Tikhonov stabilizer can be written in the quadratic form $\Omega_{TIK}(\theta) = \theta^T S \theta$ with

$$S \quad = \quad \sum_{i=1}^{n}\sum_{j=1}^{n} \int_{x \in X} \frac{\partial^2}{\partial x_i \partial x_j} \phi(x)\phi^T(x) dx \qquad (4)$$

where $\phi(x) = \nabla_\theta f(x;\theta)$ is the regressor in the linear regression formulation of the identification problem. Hence, the minimization problem remains quadratic. In the more general case when f is non-linearly parameterized, the stabilizer is no longer quadratic.

The Tikhonov stabilizer has at least two appealing interpretations that are relevant for the identification of fuzzy nonlinear models. First, it favors models that are as smooth functions of the input x 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 problem, see also the examples above. Second, it attracts the parameters towards the sub-manifold or sub-set of the parameter space that corresponds to all linear models, since the Hessian of the model is zero everywhere for a linear model. Hence, when there are excessive degrees of freedom, the model will be attracted locally towards the linear one that is most consistent with the observations, locally. Again, this is reasonable, because the basis of the TSK fuzzy model is that the system can be accurately described with linear models locally. Of course, this is related to the smoothness properties of the system.

3.3 Approximate Tikhonov Regularization

Unfortunately, it may sometimes be difficult or time-consuming to explicitly evaluate the derivatives and integral in the stabilizer (3), either exactly or numerically. While the stabilizer (3) is very general, the structure of the TSK fuzzy model suggests a simpler alternative that is motivated by the fact that the local models are assumed to have the same (not necessarily linear) structure

$$\Omega_{\mathrm{TIK}}^{\mathrm{approx}}(\theta) = \sum_{i=1}^{m}\sum_{j=1}^{i-1}(\theta_i-\theta_j)^T W(\theta_i-\theta_j)\omega_{i,j} \quad (5)$$

where $\omega_{i,j}$ is a measure of how close the fuzzy sets X_i and X_j are, in the topology of $X \subset R^N$. Hence, $\omega_{i,j}$ is close to one for neighboring fuzzy sets X_i and X_j and close to zero for a distant pair of fuzzy sets. The matrix W is a positive definite diagonal weighting matrix that ensures that the different local model parameters have sensible relative weights in the stabilizer. The interpretation of this approximation is that the local model parameters θ_i and θ_j corresponding to neighboring regions X_i and X_j should be similar, while there should not be any such constraint on local model parameters of regions that are far apart. This can obviously be interpreted as smoothness, too, but simplifications are achieved because in (5) the smoothness is described in the finite-dimensional parameter space rather than the infinite dimensional function space in which f belongs in (3).

The exact definition of the weighting factor $\omega_{i,j}$ should depend on the parameterization of the membership functions for the fuzzy sets $X_1, ..., X_n$. Let us give an example of how it may be defined, using tensor product Gaussian membership functions:

$$\omega_{i,j} = \exp\left(-\frac{1}{2}\sum_{k=1}^{n}\frac{(c_{i,k}-c_{j,k})^2}{\sigma_k^2}\right)$$

where $c_i = (c_{i,1}, ..., c_{i,n})$ is the center point of the fuzzy set X_i, and σ_k is the average "standard deviation" of the membership functions defining the fuzzy sets $X_1, ..., X_m$ along the kth dimension.

4 The Bias/Variance Tradeoff

The mean squared error (MSE) of an identified model is defined as

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

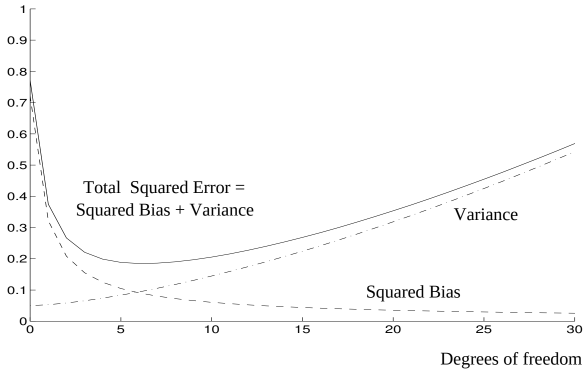

Figure 3. Typical relationship between total squared error, squared bias, variance and degrees of freedom in the model.

where the expectation E is with respect to the joint distribution of the identification data Z^N and the future input/output pair (x, y), which is assumed to be independent of Z^N. Moreover, we assume the data are generated by a system

$$y(t) = g(x(t)) + e(t)$$

where g is a function and $e(t)$ is a stochastic variable with variance λ_0. The MSE can now be decomposed into bias and variance using standard arguments [4]. With a regularized least squares method we get

$$\mathrm{MSE}_{\mathrm{REG},\gamma} = b^2(\gamma) + \left(1 + \frac{d(\gamma)}{N}\right)\lambda_0 \quad (6)$$

where $b^2(\gamma)$ is the squared bias, $d(\gamma)\lambda_0/N$ is the variance, and the degrees of freedom in the model is given by $d(\gamma)$, see [9] for full details.

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. 3. Hence, we have the well known tradeoff between bias and variance, which suggests that there exists a γ that gives model with an optimal degrees of freedom within the set of model structures. A major problem is to select γ such that it gives the optimal balance between bias and variance.

We now make the following general observation, see [9] for details: Regularization will always reduce the variance of the parameter estimate at the possible cost of a bias. The effect of the bias will depend on the prior knowledge coded into the stabilizer Ω.

The bias/variance trade-off interpretation illustrates how regularization can be applied to find the model with the best estimated 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. In other cases it may be natural to impose more smoothness constraints than what is optimal from a prediction point to views, see the second example above (Figure 2). Methods for estimating the MSE include the Final Prediction Error criterion for Regularized models (FPER) [13]

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

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

$$S(\gamma) = \left(\nabla_\theta^2 V(\hat{\theta}_\gamma(Z^N); Z^N) + \gamma \nabla_\theta^2 \Omega(\hat{\theta}_\gamma(Z^N))\right)^{-1} \cdot \nabla_\theta^2 V(\hat{\theta}_\gamma(Z^N); Z^N)$$

This criterion is a generalization of the well known Final Prediction Error (FPE) criterion, e.g. [21]. One way to choose the regularization parameter γ that minimizes FPER:

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

which can be found using a simple line search algorithm. Alternative approaches includes the use of a separate data set, data re-sampling techniques like cross-validation or generalized cross-validation [2] to estimate the MSE.

5 Simulation Example

This simulation example is based on the results reported in [3], 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:

$$\begin{aligned}
\dot{x}_1 &= \mu_m \frac{x_1 x_4 x_5}{K_s x_5 + K_0 x_4 + x_4 x_5} \\
\dot{x}_2 &= v_L \frac{x_1 x_4}{K_L + x_4} - 0.9082 K_p x_2 \\
\dot{x}_3 &= K_p x_2 \\
\dot{x}_4 &= -\frac{1}{Y_s} \mu_m \frac{x_1 x_4 x_5}{K_s x_5 + K_0 x_4 + x_4 x_5} \\
&\quad -1.011 v_L \frac{x_1 x_4}{K_L + x_4} \\
\dot{x}_5 &= k_l a(x_5^* - x_5) - 0.09 v_L \frac{x_1 x_4}{K_L + x_4} \\
&\quad -\frac{1}{Y_0} \mu_m \frac{x_1 x_4 x_5}{K_s x_5 + K_0 x_4 + x_4 x_5}
\end{aligned}$$

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 temperature and pH. This dependency is given by an interpolated lookup table based on the experimental data. Initial values for the batch are $x_1(0) = x_{10}, x_2(0) = 0, x_3(0) = 0, x_4(0) = x_{40}$, and $x_5(0)$ is determined on the basis of the other initial values.

The model structure developed in [3] is a TSK fuzzy model using local linear models within each region of the fuzzy partitioning. All the local models are chosen to have the same linear structure

$$x(t+1) = a_i + A_i x(t) + B_i u(t) \tag{7}$$

where $x = (x_1, ..., x_5)$ and $u = (u_1, u_2)$. By qualitatively examining the main reaction mechanisms, 16 regions and local models are selected, see [3] for details.

In [3] 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 from the intervals $[0.4, 0.5]$ and $[40, 50]$, respectively. 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 will rapidly become 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 are very sparse data, while the TSK fuzzy model structure contains local models 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, the FPER criterion was used to select the regularization parameter $\gamma = \hat{\gamma}(Z^N)$.

1. **Standard least squares criterion without regularization.**

2. **Least squares criterion with Tikhonov regularization,** cf. section 3.2. In the implementation the derivatives and integral in (4) is computed using finite-difference numerical approximations. The uniform grid contains about 280.000 points.

3. **Least squares with Levenberg-Marquardt regularization.** In particular, $\theta^{\#} = 0$.

4. **Least squares solved with PCA regularization**. Singular value decomposition is used to invert the Hessian matrix in the least squares problem. Singular values less than a threshold are set to infinity. This threshold is the regularization parameter.

The models will be "validated" on a number of independent 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 [3]. The average prediction error on the separate test data set is illustrated in Figure 4a 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 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 4c. 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 and more ill-conditioned. We also observe significant differences between the regularization methods, which are related to the different levels of bias introduced by the different regularization methods. We think it is characteristic that Tikhonov regularization introduce less bias than for instance PCA based regularization, 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. Another test of the identified model is provided in Figure 4b, which illustrates the control performance in terms of average production rate of gluconic acid with optimal controllers designed on the basis of the identified

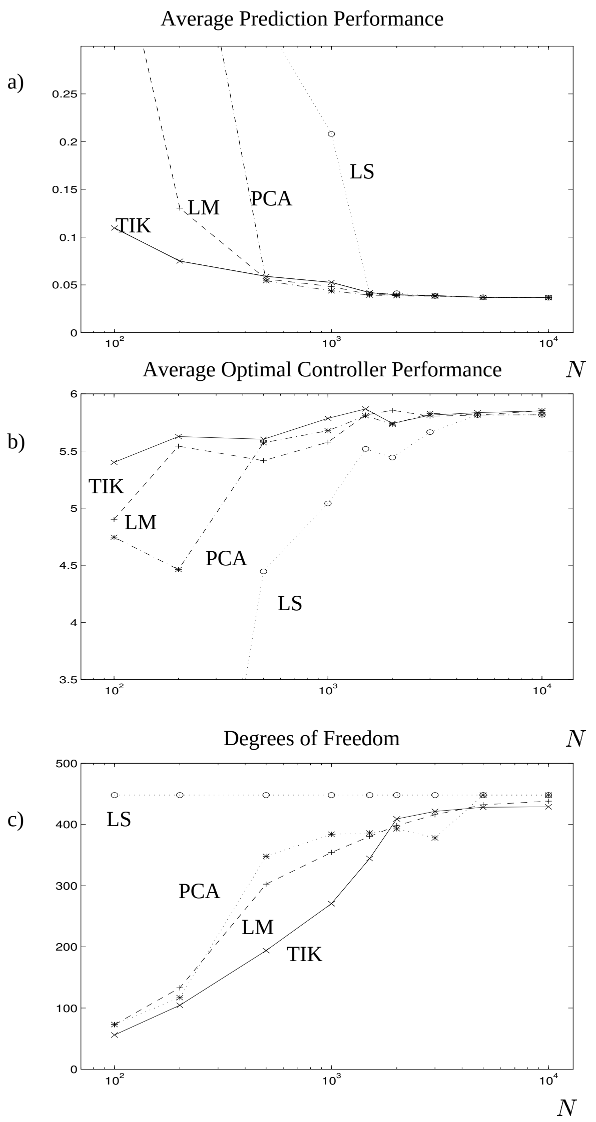

Figure 4. Results of simulation experiments: a) Illustrates the estimated average prediction performance as a function of the number of observations N. b) Illustrates the optimal control performance in terms of average production rate of gluconic acid as a function of the number of observations N. c) 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 PCA is marked with $\star$, models identified using Levenberg-Marquardt regularization is marked with $+$, while models identified with Tikhonov regularization is marked with $\times$.

models. This also indicates the same rankings of the identified models.

6 Concluding Remarks

In practical applications, with highly complex fuzzy non-linear models, non-uniform distribution of data, and possibly over-parameterized models, problems should be expected when identifying fuzzy models. One 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 introducing priors such as smoothness.

A major benefit of regularization is that the sensitivity of the identified model to the choice of fuzzy model structure will be reduced. For example, it is perfectly safe to use a uniform partitioning in the fuzzy model, provided regularization is applied. Without regularization, however, it is our experience that such a choice rarely leads to an accurate and robust model. Hence, we believe that if one do not want to (or is unable to) invest sufficient effort in finding a parsimonious fuzzy model structure, regularization is a valuable tool that may strongly improve the robustness of the identification algorithm and eventually give a better model. Related work that discuss the use of regularization in identification of fuzzy (or closely related) models can be found in [9, 8, 16].

References

[1] C. Bishop. Improving the generalization properties of radial basis function neural networks. *Neural Computation*, 3:579–588, 1991.

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

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

[4] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

[5] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural network architectures. *Neural Computation*, 7:219–269, 1995.

[6] R. J. Hathaway and J. C. Bezdek. Switching regression models and fuzzy clustering. *IEEE Trans. Fuzzy Systems*, 1:195–204, 1993.

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

[8] T. A. Johansen. Identification of non-linear systems using empirical data and prior knowledge - an optimization approach. *Automatica*, 32:337–356, 1996.

[9] T. A. Johansen. On Tikhonov regularization, bias and variance in nonlinear system identification. To appear in Automatica, 1996.

[10] T. A. Johansen and B. A. Foss. Identification of nonlinear system structure and parameters using regime decomposition. *Automatica*, 31:321–326, 1995.

[11] T. Kavli. ASMOD – An algorithm for adaptive spline modelling of observation data. *Int. J. Control*, 58:947–967, 1993.

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

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

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

[15] D. W. Marquardt. An algorithm for least squares estimation of nonlinear parameters. *J. SIAM*, 11:431–441, 1963.

[16] R. Murray-Smith and T. A. Johansen. Local learning for local model networks. In *Proc. 4th IEE Int. Conference on Artificial Neural Networks, Cambridge, UK*, pages 40–46, 1995.

[17] R. Murray-Smith and T. A. Johansen, editors. *Multiple Model Approaches to Modelling and Control.* Taylor and Francis, London, 1996.

[18] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1:502–527, 1986.

[19] T. Poggio, V. Torr, and C. Koch. Computational vision and regularization. *Nature*, 317:314–319, 1985.

[20] J. Sjöberg, T. McKelvey, and L. Ljung. On the use of regularization in system identification. In *Preprints 12th IFAC World Congress, Sydney*, volume 7, pages 381–386, 1993.

[21] T. Söderström and P. Stoica. *System Identification.* Prentice Hall, Englewood Cliffs, NJ, 1988.

[22] M. Sugeno and G. T. Kang. Structure identification of fuzzy model. *Fuzzy Sets and Systems*, 26:15–33, 1988.

[23] B. F. Swindel. Good ridge estimates based on prior information. *Comm. Statistics*, A5:985–997, 1976.

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

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

[26] G. Wahba. *Spline Models for Observational Data.* SIAM, Philadelphia, 1990.

[27] R. R. Yager and D. P. Filev. Unified structure and parameter identification of fuzzy models. *IEEE Trans. Systems, Man, and Cybernetics*, 23:1198–1205, 1993.