Gestel T., Suykens J.A.K., Van Dooren P., De Moor B., "Imposing stability in subspace identification by regularization", in *Proc. of the 39th IEEE Conference on Decision and Control (CDC2000)*, Sydney, Australia, Dec. 2000, pp. 1555-1560.,

# Imposing Stability in Subspace Identification by Regularization

Tony Van Gestel[1], Johan A.K. Suykens[1],
Paul Van Dooren[2] and Bart De Moor[1]

**Abstract** –In subspace methods for linear system identification, the system matrices are usually estimated by least squares, based on estimated Kalman filter state sequences and the observed inputs and outputs. For an infinite number of data points and a correct choice of the system order, this least squares estimate of the system matrices is unbiased. However, when using subspace identification on a finite number of data points, the estimated model can become unstable, for a given deterministic system which is known to be stable. In this paper, stability of the estimated model is imposed by adding a regularization term to the least squares cost function. The regularization term used here is the trace of a matrix which involves the dynamical system matrix and a positive (semi-)definite weighting matrix. The amount of regularization needed can be determined by solving a generalized eigenvalue problem. It is shown that the so-called data augmentation method proposed by Chui and Maciejowski corresponds to adding regularization terms with specific choices for the weighting matrix. The choice of the identity matrix for the weighting matrix is motivated by simulation results.

**Keywords:** stability, subspace identification, regularization, ridge regression

## 1 Introduction

The linear combined deterministic-stochastic identification problem is concerned with systems and models of the form:

$$x_{k+1} = Ax_k + Bu_k + w_k \quad (1)$$
$$y_k = Cx_k + Du_k + v_k \quad (2)$$

with[1]

$$\mathbf{E}\left( \begin{bmatrix} w_p \\ v_p \end{bmatrix} \begin{bmatrix} w_p \\ v_p \end{bmatrix}^T \right) = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \delta_{pq} \geq 0. \quad (3)$$

The vectors $u_k \in \mathbb{R}^m$ and $y_k \in \mathbb{R}^l$ with discrete time index $k$ denote the $m$ inputs and $l$ outputs of the system respectively. The $n$ states at the time index $k$ of the system with order $n$ are denoted by the state vector $x_k \in \mathbb{R}^n$. The process noise $w_k \in \mathbb{R}^n$ and the measurement noise $v_k \in \mathbb{R}^l$ are assumed to be zero mean, white Gaussian with covariance matrices as given by (3). The model matrices $A, B, C, D$ and the covariance matrices $Q, S, R$ have appropriate dimensions. Both the deterministic and stochastic identification problem are special cases of the combined identification problem, without noise inputs $w_k = 0$ and $v_k = 0$ in (1)-(3) and no deterministic inputs $u_k = 0$, respectively.

The system identification problem is stated as follows: *given* a large number of observations of the input $u_k$ and the corresponding output $y_k$ generated by the unknown system (1)-(3), *determine* the system order $\hat{n}$ of the unknown system (1)-(3) and the estimates $\hat{A}, \hat{B}, \hat{C}, \hat{D}$ up to within a similarity transformation together with the estimated noise covariances matrices $\hat{Q}, \hat{S}, \hat{R}$ so that the second order statistics of the output of the model *and* of the given output are equal (or equivalent in the sense of Faurre [8]).

In the last decade, so-called subspace identification methods [8] have been developed. Typically, in a first step, Kalman filter state sequences $\hat{X}_i \in \mathbb{R}^{\hat{n} \times j}$ and $\hat{X}_{i+1} \in \mathbb{R}^{\hat{n} \times j}$ of the system are estimated directly from input-output data using geometric operations of subspaces spanned by the column or row vectors of block Hankel matrices formed by input-output data. The output data block Hankel matrix $Y_{0|i-1}$ is constructed from the observations $y_0, y_1, \ldots, y_{i+j-2}$ as follows:

$$Y_{0|i-1} = \begin{bmatrix} y_0 & y_1 & \cdots & y_{j-1} \\ y_1 & y_2 & \cdots & y_j \\ \cdots & \cdots & \cdots & \cdots \\ y_{i-1} & y_i & \cdots & y_{i+j-2} \end{bmatrix},$$

where $i$ is a user defined dimension with $i \ll j$ and the subscript $0|i-1$ denotes the time indices of the

[1]Katholieke Universiteit Leuven, Dept. of Electrical Engineering ESAT-SISTA, Kard. Mercierlaan 94, B-3001 Leuven, Belgium. T. Van Gestel, J.A.K Suykens and B. De Moor are a Research Assistant, a Postdoctoral Researcher and a Research Associate with the Fund for Scientific Research-Flanders (FWO-Vlaanderen), respectively. Email: {tony.vangestel,johan.suykens,bart.demoor}@esat.kuleuven.ac.be. Tel. +32 (16) 32.17.09. Fax. +32 (16) 32.19.70.

[2]P. Van Dooren is Professor at the Catholic University of Louvain, Dept. of Mathematical Engineering, Av. Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium. Email: vdooren@csam.ucl.ac.be. Tel. +32 (10) 47.80.40. Fax. +32 (10) 47.21.80

[1]$\mathbf{E}$ denotes the expected value operator and $\delta_{pq}$ the Kronecker delta. It is assumed that the process is stationary and ergodic: $\mathbf{E}[a_k b_k^T] = \lim_{j \to \infty} [\frac{1}{j} \sum_{i=0}^{j} a_i b_i^T]$.

outputs in the first column of $Y_{0|i-1}$. The input data block Hankel matrix $U_{0|i-1}$ is constructed in a similar way. The calculations are performed in a numerically reliable way, based on the singular value decomposition (SVD) and QR-decomposition (see [8] for details).

After the estimation of the Kalman filter states sequences $\hat{X}_i$ and $\hat{X}_{i+1}$ in the first step, the system matrices $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ of the combined stochastic-deterministic identification problem are identified in the second step:

$$\min_{\hat{A},\hat{B},\hat{C},\hat{D}} \left\| \begin{bmatrix} \hat{X}_{i+1} \\ Y_{i|i} \end{bmatrix} - \begin{bmatrix} \hat{A} & \hat{B} \\ \hat{C} & \hat{D} \end{bmatrix} \cdot \begin{bmatrix} \hat{X}_i \\ U_{i|i} \end{bmatrix} \right\|_F^2 . \quad (4)$$

Motivated by consistency results [8] for $j \to \infty$, this optimization problem is solved in a least squares sense. However, when identifying on a finite number of observations (finite $j$), the least squares estimate does not always yield a stable system [2], while often the real system is known to be stable. For a finite number of data points, this may e.g. occur in the presence of high noise levels or when overparameterizing the system by choosing a too high system order. A discrete time linear system is called stable when it has all its poles inside the unit disk or when there exists a positive definite matrix $\hat{P} = \hat{P}^T > 0$ of appropriate dimension such that $\hat{A}\hat{P}\hat{A}^T - \hat{P} < 0$. In this paper, stability of the system is imposed by adding a regularization term to (4). Usually, regularization is done by adding the norm $\|\theta\|_2$ of the parameter vector $\theta$ to the cost function $J(\theta)$ of the optimization problem [1, 3, 6]. The resulting $\theta$ is then obtained as the solution of the minimization problem $\min_\theta J(\theta) + c\|\theta\|_2$. In this paper, a new and specific type of regularization term is used: stability is obtained by adding the trace of product of the system matrix, a positive semi-definite matrix $W \geq 0$, and the transpose of the system matrix, $\mathrm{Tr}(\hat{A}W\hat{A}^T)$, to the cost function (4). Upper bounds for the spectral radius are obtained and it is shown that the calculation of the amount of regularization needed to obtain a specific spectral radius boils down to solving a Generalized Eigenvalue problem.

Also for the stochastic identification problem, the least squares estimate

$$\min_{\hat{A},\hat{C}} \left\| \begin{bmatrix} \hat{X}_{i+1} \\ Y_{i|i} \end{bmatrix} - \begin{bmatrix} \hat{A} \\ \hat{C} \end{bmatrix} \cdot \hat{X}_i \right\|_F^2 \quad (5)$$

does not always yield a stable system. Since the stochastic identification least squares estimate (5) is a special case of the combined estimate (4) with $U_{i|i} = 0$, in the sequel of the paper, the more general case of stability of (4) is discussed.

This paper is organized as follows. In Section 2, the use of regularization to impose stability is explained. In Section 3, the use of regularization is motivated by simulation results. This paper is a companion paper of [7], which contains all the proofs. Instead, this paper focuses on the simulation results.

The eigenvalues and singular values of a square matrix $A \in \mathbb{R}^{n\times n}$ are denoted by $\lambda_i(A)$ and $\sigma_i(A)$ respectively, for $i = 1, \ldots, n$. The spectral radius of $A$ is denoted by $\rho(A) = \max_{i=1,\ldots,n} |\lambda_i(A)|$. The minimal and maximal singular value of $A$ is denoted by $\underline{\sigma}(A)$ and $\overline{\sigma}(A)$, respectively. The trace of the matrix $A$ is denoted by $\mathrm{Tr}(A)$ and the matrix norm $\overline{\sigma}(MAM^{-1})$, with $M \in \mathbb{R}^{n\times n}$ a non-singular matrix of appropriate dimensions, is denoted by $\|A\|_M$. The Kronecker product of two matrices $X$ and $Z$ is denoted by $X \otimes Z$.

## 2 Stability by using Regularization

The estimation problem that we consider is the following: given the matrices $\hat{X}_{i+1}$, $Y_{i|i}$, $\hat{X}_i$ and $U_{i|i}$ from the first step, estimate the model $\hat{A}$, $\hat{B}$, $\hat{C}$, $\hat{D}$. The least squares estimates (4) and (5) do not guarantee a stable $\hat{A}$ matrix for given finite data, while often the true linear model (1)-(3) is known to be stable.

The least squares solution of (4) for $[\hat{A}\ \hat{B}]$ is given by:

$$\begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix} = \hat{X}_{i+1} \cdot \begin{bmatrix} \hat{X}_i^T & U_{i|i}^T \end{bmatrix} \cdot \begin{bmatrix} \hat{X}_i\hat{X}_i^T & \hat{X}_iU_{i|i}^T \\ U_{i|i}\hat{X}_i^T & U_{i|i}^TU_{i|i} \end{bmatrix}^{-1} . \quad (6)$$

For a finite number of data points, the estimates of $\hat{X}_i$ and $\hat{X}_{i+1}$ may result in an unstable estimate of the system matrix $\hat{A}$ in (6).

Stability of the model can be imposed by adding a regularization term to the cost function from which (6) is determined, i.e.:

$$\min_{\tilde{A}_c,\tilde{B}} J(\tilde{A}_c, \tilde{B}) = J_1(\tilde{A}_c, \tilde{B}) + cJ_2(\tilde{A}_c), \quad (7)$$

with

$$J_1(\tilde{A}_C, \tilde{B}) = \left\| \hat{X}_{i+1} - [\tilde{A}_c\ \tilde{B}] \cdot [\hat{X}_i^T\ U_{i|i}^T]^T \right\|_F^2 , \quad (8)$$

$$J_2(\tilde{A}_c) = \|\tilde{A}_cQ\|_F^2 = \mathrm{Tr}(\tilde{A}_cW\tilde{A}_c^T). \quad (9)$$

The amount of regularization is characterized by the positive real scalar $c$ and by the positive semidefinite matrix $W = QQ^T \geq 0$. The optimal solution to (7) is then given by[2]:

$$[\tilde{A}_c\ \tilde{B}] = [\hat{A}\ \hat{B}]\hat{\Sigma}_{XU}(\hat{\Sigma}_{XU} + cW_e)^{-1}, \quad (10)$$

with

$$\hat{\Sigma}_{XU} = \begin{bmatrix} \hat{X}_i\hat{X}_i^T & \hat{X}_iU_{i|i}^T \\ U_{i|i}\hat{X}_i^T & U_{i|i}U_{i|i}^T \end{bmatrix}, \quad W_e = \begin{bmatrix} W & 0_{\hat{n}\times m} \\ 0_{m\times\hat{n}} & 0_{m\times m} \end{bmatrix} . \quad (11)$$

[2] In the sequel of this paper, estimates from (6) are denoted by $\hat{A}$ and estimates from (7) by $\tilde{A}$.

In [3], regularization (ridge regression) is used to obtain a reduction in the variance of the estimate, while allowing for a (small) bias. The regularization parameter $c$ is chosen in such a way that a cross-validation weighted square error is minimized. One typically uses $W = I_{\hat{n}}$, motivated by [4]. In this paper, the regularization parameter $c$ is chosen such that a stable system matrix is obtained, i.e. $\rho(\tilde{A}_c) < 1$ or such that $\tilde{A}_c$ has a spectral radius $\rho(\tilde{A}_c)$ smaller than $\gamma$, i.e., $\tilde{A}_c \tilde{P} \tilde{A}_c^T - \gamma^2 \tilde{P} < 0$, with $\tilde{P} = \tilde{P}^T > 0$ and $\gamma$ a positive real constant.

The use of regularization to obtain a stable system matrix estimate $\tilde{A}_c$ is motivated by the following inequality:

$$\begin{aligned} \|\tilde{A}_c W^{\frac{1}{2}}\|_F^2 &= \mathrm{Tr}(\tilde{A}_c W \tilde{A}_c^T) = \textstyle\sum_{i=1}^n [\sigma_i(\tilde{A}_c W^{\frac{1}{2}})]^2 \\ &\geq \textstyle\sum_{i=1}^n \underline{\sigma}(W)[\sigma_i(\tilde{A}_c)]^2 \geq \underline{\sigma}(W) \sum_{i=1}^n |\lambda_i(\tilde{A}_c)|^2. \end{aligned}$$

Let $\tilde{A}_c = \hat{A}\hat{\Sigma}_{XU}(\hat{\Sigma}_{XU} + cW_e)^{-1}$ denote the solution $\tilde{A}_c$ of (7) for a given $c$, it follows from the optimality of the least squares estimate (10) that $J_2(\tilde{A}_{c_2}) \leq J_2(\tilde{A}_{c_1})$ for $c_1 \leq c_2$, i.e., the regularization term evaluated at the corresponding solutions for $c$ is a non-increasing function of $c$. However, the spectral radius $\rho(\tilde{A}_c)$ is not always a monotonically decreasing function for increasing $c$. In the sequel of this Section, this behavior is characterized in more detail by deriving an upper bound for $\rho(\tilde{A}_c)$ and by calculating the amount of regularization needed to obtain $\rho(\tilde{A}_c) = \gamma < 1$. We refer to [7] for all the proofs.

The relation between $\hat{A}$ and $\tilde{A}_c$ is given by (10). By defining

$$\tilde{A}_e = \begin{bmatrix} \tilde{A}_c & \tilde{B} \\ 0_{m\times\hat{n}} & 0_{m\times m} \end{bmatrix}, \quad \hat{A}_e = \begin{bmatrix} \hat{A} & \hat{B} \\ 0_{m\times\hat{n}} & 0_{m\times m} \end{bmatrix}, \quad (12)$$

we have that the eigenvalues $\lambda(\tilde{A}_c)$ of $\tilde{A}_c$ are equal to the $\hat{n}$ eigenvalues of largest modulus of the (extended) square matrix $\tilde{A}_e$. Given (10), the influence of regularization on the eigenvalues of $\tilde{A}$ is given by the relation $\tilde{A}_e = \hat{A}_e\hat{\Sigma}_{XU}(\hat{\Sigma}_{XU} + cW_e)^{-1}$. However, the increase in dimensionality from $\hat{n}$ to $\hat{n} + m$ can be avoided:

**Lemma 1** *Let $\hat{\Sigma}_s$ be defined as*

$$\hat{\Sigma}_s = \hat{X}_i\hat{X}_i^T - \hat{X}_i U_{i|i}^T (U_{i|i}U_{i|i}^T)^{-1} U_{i|i}\hat{X}_i^T, \quad (13)$$

*then the eigenvalues of $\tilde{A}$ are equal to*

$$\lambda(\tilde{A}_c) = \lambda(\hat{A}_c\hat{\Sigma}_s(\hat{\Sigma}_s + cW)^{-1}). \quad (14)$$

By using the QR-decomposition

$$\left[\begin{array}{c|c} U_{i|i}^T & \hat{X}_i^T \end{array}\right] = \left[\begin{array}{c|c} Q_1 & Q_2 \end{array}\right] \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix},$$

it is easily shown that $\hat{\Sigma}_s = R_{22}^T R_{22}$ or $\hat{\Sigma}_s = \hat{X}_s\hat{X}_s^T$, with $\hat{X}_s = \hat{X}_i - \hat{X}_i U_{i|i}^T (U_{i|i}U_{i|i}^T)^{-1} U_{i|i}$. An upper bound for the spectral radius $\rho(\tilde{A}_c)$ can be formulated as follows:

**Theorem 1** *Let $W > 0$, then the following upper bound holds for $\rho(\tilde{A}_c)$:*

$$\rho(\tilde{A}_c) \leq \frac{\overline{\sigma}(R_{22}^{-T}\hat{A}R_{22}^T)}{1 + c\,\underline{\sigma}(R_{22}^{-T}WR_{22}^{-1})}. \quad (15)$$

*From this upper bound, it follows that*

$$\rho(\tilde{A}_c) \leq \gamma \quad \textit{for} \quad c \geq \frac{\left(\frac{\overline{\sigma}(R_{22}^{-T}\hat{A}R_{22}^T)}{\gamma} - 1\right)}{\underline{\sigma}(R_{22}^{-T}WR_{22}^{-1})}.$$

This Theorem tells us that the upper bound (15) for the spectral radius $\rho(\tilde{A}_c)$ is a monotonically decreasing function for increasing $c$. Hence, we can always find a $c$ to impose stability.

In the next Theorem, conditions are derived to obtain $\rho(\tilde{A}_c) \leq \gamma$, with $\gamma > 0$ a positive real scalar.

**Theorem 2** *Given $\hat{X}_{i+1}$, $\hat{X}_i$, $U_{i|i}$, $Y_{i|i}$ and $W$. Let the matrix $\hat{\Sigma}_s$ be defined by (13). Let $\gamma \geq 0$ be a positive real scalar and let $\hat{A}_c$ and $\tilde{A}_c$ be estimated as in (4) and (10). Let the matrices $P_2$, $P_1$ and $P_0 \in \mathbb{R}^{\hat{n}\times\hat{n}}$ be defined as follows: $P_2 = -\gamma W \otimes \gamma W$, $P_1 = -\gamma^2 W \otimes \hat{\Sigma}_s - \gamma^2\hat{\Sigma}_s \otimes W$, $P_0 = \hat{A}\hat{\Sigma}_s \otimes \hat{A}\hat{\Sigma}_s - \gamma\hat{\Sigma}_s \otimes \gamma\hat{\Sigma}_s$. Define the set of eigenvalues $\vartheta$ of the Generalized Eigenvalue problem*

$$\vartheta = \lambda\left(\begin{bmatrix} 0 & -I \\ P_0 & P_1 \end{bmatrix}, -\begin{bmatrix} I & 0 \\ 0 & P_2 \end{bmatrix}\right). \quad (16)$$

*Then $\rho(\tilde{A}_c) \leq \gamma$, for $c \geq c_m = \max_{i|\vartheta_i\in\mathbb{R}^+}\vartheta_i$, with $\rho(\tilde{A}_{c_m}) = \gamma$.*

For high order systems, a large Generalized Eigenvalue Problem (16) of dimension $2\hat{n}^2$ has to be solved, hence requiring $O(\hat{n}^6)$ operations. When $\hat{n}$ is large, an iterative algorithm to determine the value of $c_m$ is proposed in [7]. The algorithm requires $O(\hat{n}^3)$ operations per iteration step and the simulation results indicate linear convergence. The problem can also be formulated as a real stability radius problem with one repeated block for which a fast algorithm exists [5].

Now, we show that the so-called data augmentation method of Chui and Maciejowski [2] corresponds to adding regularization terms to the least squares cost function with specific choices for the weighting matrix $W$. In the method of [2], the non-steady Kalman filter state sequences $\hat{X}_i, \hat{X}_{i+1}$ are iteratively augmented by appending $\sqrt{c_p}V_p$ and $0_{\hat{n}\times 1}$ ($\sqrt{c_q}[V_q\ V_q^*]$ and $0_{\hat{n}\times 2}$) respectively, for each unstable pole $\lambda_p$ (pole pair $(\lambda_q, \lambda_q^*)$) with corresponding normalized right eigenvector $V_p$ (right eigenvector $V_q$ and its conjugate $V_q^*$). The inputs $U_{i|i}$ and outputs $Y_{i|i}$ are extended by appending null vectors of appropriate dimensions. The constant $c_p$ ( $c_q$ for a complex pole pair) is determined such that

the magnitude of the stabilized pole is $M$, where $M \geq 0$ is chosen by the user. The other eigenvalues are not changed.

For the case of one real unstable pole, this method corresponds to minimizing the following cost function in least squares sense:

$$\min_{\tilde{A}_{CM},\tilde{B}_{CM}} \|\hat{X}_{i+1} - \tilde{A}_{CM}\hat{X}_i - \tilde{B}_{CM}U_{i|i}\|_F^2 + c_p\|\tilde{A}_{CM}V\|_F^2,$$

with the regularization term $c_p\|\tilde{A}_{CM}V\|_F^2 = c_{CM}\mathrm{Tr}(\tilde{A}_{CM}VV^T\tilde{A}_{CM}^T)$. More regularization terms are iteratively added for each other pole or pole pair with amplitude larger than $\gamma$. Hence, the data augmentation method of Chui and Maciejowski [2] corresponds to the following weighting matrix in the regularization term (7):

$$c\mathrm{Tr}(\tilde{A}_cW\tilde{A}_c^T) = \sum_{\text{unstable poles } p} c_p\mathrm{Tr}(\tilde{A}_cV_pV_p^T\tilde{A}_c^T) + \sum_{\text{unstable pole pairs } q} c_q\mathrm{Tr}(\tilde{A}_c[V_q\ V_q^*][V_q\ V_q^*]^T\tilde{A}_c^T),$$

or $cW = \sum_p c_pV_pV_p^T + \sum_q c_q[V_q\ V_q^*][V_q\ V_q^*]^T$, which can be considered as a special case of $cJ_2(\tilde{A})$, with $J_2(\tilde{A})$ from (9).

## 3 Simulation results

In this Section, the use of regularization to impose stability of the estimated system is illustrated by means of two examples. In the first example, the original system has order $n = 3$ and the effects of high noise levels for both process and measurement noise, over-estimating the system order and a small number of given data points are illustrated. In the second example, a model obtained from data of an industrial dryer [2] with order $n = 10$ is used.

For the first example, the original system is given by:

$$A = \begin{bmatrix} -0.5 & 0 & 0 \\ 0 & 0.65 & 0.70 \\ 0 & -0.70 & 0.65 \end{bmatrix},\ B = \begin{bmatrix} 10 \\ 1 \\ 1 \end{bmatrix},\ C = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}^T, \quad (17)$$

with $D = 1$. The known input $u_k$ has a standard normal distribution. The influence of the noise level of covariance matrices $Q$, $R$, $S$ (3) is illustrated. For the weighting matrix $W$, the identity matrix $W = I_{\hat{n}}$ is chosen [1, 3, 4]. The evolution of the spectral radius $\rho(\tilde{A}_c)$ and the different parts of the cost function (7) as a function of $c$ is depicted in Figure 1 for an estimated unstable system with system order $\hat{n} = 3$.

Monte Carlo simulations were used to compare the performance of the following 3 identification methods: a) N4SID [8] with full (`SI`) and with reduced system order (`SIr`), where the system order was reduced until a stable estimate was obtained; b) (full order) subspace identification with regularization in the sense of Chui and Maciejowski (`CM`); c) (full order) subspace identification with regularization with $W = I_{\hat{n}}$ (`Reg`). The following performance criteria were used:

1. the difference between the impulse response (IR) of the true deterministic system (17) and of the estimated deterministic system: $((D - \hat{D})^2 + \sum_{i=1}^{k-1}(CA^{i-1}B - \hat{C}\tilde{A}^{i-1}\tilde{B})^2)^{\frac{1}{2}}$, for $k = 10$ and $k = 50$ respectively.

2. mean(var($\bar{A}_{ij}$)): the mean of the variances the elements of estimated system matrix $\bar{A}$, where the variance is taken over all Monte Carlo simulations and the mean over all elements of $\bar{A}$. This criterion is used to illustrate the effect of regularization on the variance of the estimates[3] [1, 3, 6]. The estimate $\bar{A}$ corresponds to $\hat{A}$, $\tilde{A}_{CM}$ and $\tilde{A}$ for the estimates obtained by N4SID (`SI`) [8], the methods `CM` and `Reg`, respectively.

The effect of the finite number of data points (finite $j$), of a too high system order (or overparameterization) and of high noise levels on a finite number of data points are summarized in Table 1. For each experiment 100 unstable (full order model) estimates were generated. For each simulation, the percentage of estimated models with order $\hat{n}$ that were unstable is reported. Both regularization methods yield better results than the `SIr` method. For high noise levels and for a small number of data points, `Reg` performs better results than `CM`. Notice that the `Reg` method reduces the variance on the estimate $\tilde{A}$ significantly, as is expected from regularization theory [1, 4, 3, 6]. We see the influence of the regularization on the distribution of the poles illustrated in Figure 2.

In the second example, a (stable) 10th order model was used. The setup of the simulations corresponds to that of the first example and corresponds to the approach in [2]. Table 2 illustrates the effect of high noise levels, the number of data points and overparameterization. For 80 data points, `Reg` performs better than `CM` and `SIr`, particularly for high noise levels. The variance on the elements of $\tilde{A}$ is kept low. For 160 data points, the difference in performance between the two regularization methods `CM` and `Reg` decreases. For $\hat{n} = 12$, the `SIr` yields the best results. The main cause for instability are the 11th and 12th poles which do not correspond to a pole of the original system.

We now summarize the results in Tables 2 and 3 obtained with the methods `CM` and `Reg`, corresponding to different choices for the weighting matrix $W$. It is

[3] Note that for $j \to \infty$ all models are identified in the same state space basis.

observed that the ridge regression approach Reg results generally in a better performance than CM. The increase in performance is rather small when compared to the results of both methods CM and Reg with the SIr method. Together with imposing stability, the Reg method with $W = I_{\hat{n}}$ significantly reduces the variance of the estimate, especially for overparameterized systems. Recall that regularization is typically used [1, 4, 3, 6] to reduce the variance. Especially for overparameterized systems, the weighting matrix $W$ of the CM method is too specific [4] and results in a smaller reduction of the variance.

Finally, we compare the speed up of the iterative algorithm [7] with the solving the (large) Generalized Eigenvalue problem, by counting the mean number of flops used by both methods. For $\hat{n} = 10, 12$, the iterative algorithm (with $\epsilon_{\text{tol}} = 10^{-6}$) needed $6 \times 10^6$, $11 \times 10^6$ flops respectively, while solving (16) took $660 \times 10^6$, $1950 \times 10^6$ flops respectively. The simulation results of Table 3 indicate linear convergence for the iterative algorithm [7].

## 4 Conclusions

In this paper, a regularization approach is proposed to impose stability in subspace identification when it is known that the true linear system is stable. The regularization term used here is the trace of a matrix which involves the dynamical system matrix and a positive (semi-)definite weighting matrix. The amount of regularization is calculated by solving a generalized eigenvalue problem. The so-called data augmentation method proposed by Chui and Maciejowski can be interpreted as iteratively applying regularization with specific choices for the weighting matrix. The choice of the identity matrix for the weighting matrix is motivated from results in regularization theory and by the empirical results on two datasets.

## Acknowledgments

This work was partially supported by grants and projects from the Flem. Gov.: (Res. Council K.U.Leuven: Grants, GOA-Mefisto 666; FWO: Grants, res. proj. and comm. (ICCoS and ANMMM); AWI: Bil. Int. Coll.; IWT: STWW Eureka SINOPSYS, IMPACT); from the Belgian Fed. Gov. (IUAP-IV/02, IV/24; Program Dur. Dev.); from the Eur. Comm.: (TMR Netw. (Alapedes, Niconet); Science: ERNSI); from NSF: contract CCR-97-96315. The scientific responsibility is assumed by its authors.

## References

[1] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[2] N.L.C. Chui, J.M. Maciejowski, "Realization of Stable Models with Subspace Methods", *Automatica* 32, pp. 1587–1595, 1996.

[3] G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore and London, 1989.

[4] S.F. Gull, "Bayesian Inductive Inference and Maximum Entropy", in G.J. Erickson and R. Smith (Eds.), *Maximum-Entropy and Bayesian Methods in Science and Engineering, 1*, pp. 53-74, Kluwer Academic Publishers, Dordrecht, 1988.

[5] C. Lawrence, A. Tits, P. Van Dooren, "A fast algorithm for the computation of an upper bound on the $\mu$-norm", in: *Proc. IFAC World Congress (San Francisco)*, vol. H, pp. 59–64, 1996.

[6] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P. Glorennec, H. Hjalmarsson, A. Juditsky, "Nonlinear black-box modelling in system identification: a unified overview", *Automatica* 31, pp. 1691–1724, 1995.

[7] T. Van Gestel, J. Suykens, P. Van Dooren and B. De Moor, "Identification of Stable Models in Subspace Identification by using Regularization", Accepted for publication in IEEE Trans. on Automatic Control.

[8] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems: Theory, Implementation, Applications*, Kluwer Academic Publishers, Boston/London/Dordrecht, 1996.

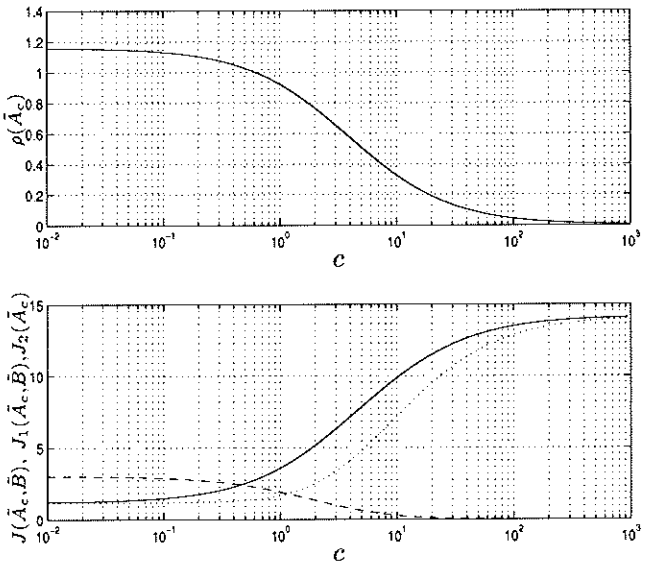


**Figure 1:** Top: evolution of $\rho(\tilde{A}_c)$ as a function of the regularization parameter $c$. Bottom: evolution of the two parts of the cost function (7) (for the optimal solution) as a function of $c$: $J_1(\tilde{A}_c, \tilde{B})$ (dashed-dotted line), $J_2(\tilde{A}_c)$ (dotted line) and the total cost function $J(\tilde{A}_c, \tilde{B})$ (full line).

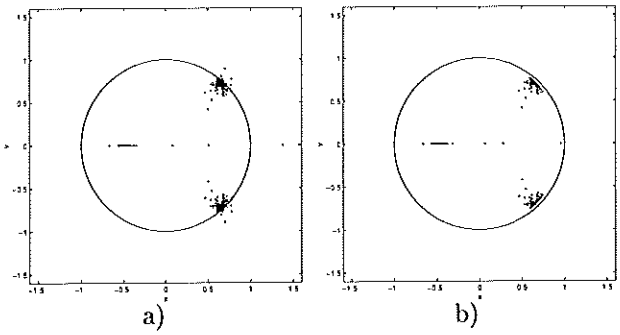

**Figure 2:** Monte Carlo simulation of the distribution of the poles $p_i = x_i + jy_i$ $(i = 1, \ldots, \hat{n})$ of the estimated system with $\hat{n} = 3$: a) no regularization, b) with regularization. The original system (17) is simulated with $Q = \text{diag}([1\ 1\ 1])$, $S = 0$, $R = 1$. The number of data points is 40, the maximal spectral radius $\rho(\tilde{A})$ is $\gamma = 0.96$.

| | IR: $k = 10$ | | | IR: $k = 50$ | | | mean(var($\tilde{A}_{ij}$)) | | |
|---|---|---|---|---|---|---|---|---|---|
| | SIr | CM | Reg | SIr | CM | Reg | SI | CM | Reg |
| $\hat{n}$=3, $Q$=diag([1 1 1]), $S$=0, $R$=1, $N_D$=40, 17% unst. | | | | | | | | | |
| Mean | 5.43 | 2.12 | 2.01 | 6.50 | 4.70 | 3.46 | 0.23 | 0.20 | 0.18 |
| Var | 5.31 | 0.60 | 0.41 | 9.61 | 3.50 | 1.46 | | | |
| $\hat{n}$=4, $Q$=diag([1 1 1]), $S$=0, $R$=1, $N_D$=40, 38% unst. | | | | | | | | | |
| Mean | 4.01 | 2.75 | 2.67 | 5.26 | 5.32 | 4.04 | 0.93 | 0.18 | 0.14 |
| Var | 6.07 | 1.28 | 0.82 | 5.71 | 7.72 | 1.96 | | | |
| $\hat{n}$=5, $Q$=diag([1 1 1]), $S$=0, $R$=1, $N_D$=40, 77% unst. | | | | | | | | | |
| Mean | 3.14 | 3.00 | 2.79 | 4.66 | 5.46 | 4.12 | 3.53 | 1.71 | 0.19 |
| Var | 3.59 | 3.93 | 0.72 | 4.21 | 8.45 | 1.23 | | | |
| $\hat{n}$=3, $Q$=diag([1 1 1]), $S$=0, $R$=1, $N_D$=80, 0.002% unst. | | | | | | | | | |
| Mean | 4.98 | 1.08 | 1.14 | 5.95 | 1.97 | 2.04 | 0.20 | 0.18 | 0.17 |
| Var | 4.36 | 0.09 | 0.08 | 3.75 | 0.50 | 0.43 | | | |
| $\hat{n}$=4, $Q$=diag([1 1 1]), $S$=0, $R$=1, $N_D$=80, 27% unst. | | | | | | | | | |
| Mean | 3.19 | 1.04 | 1.23 | 4.59 | 2.50 | 2.30 | 0.28 | 0.17 | 0.15 |
| Var | 5.08 | 0.08 | 0.13 | 5.14 | 0.62 | 0.47 | | | |
| $\hat{n}$=5, $Q$=diag([1 1 1]), $S$=0, $R$=1, $N_D$=80, 69% unst. | | | | | | | | | |
| Mean | 2.89 | 1.20 | 1.52 | 4.27 | 2.96 | 2.60 | 1.25 | 0.50 | 0.14 |
| Var | 2.27 | 0.17 | 0.25 | 2.93 | 4.11 | 0.65 | | | |
| $\hat{n}$=3, $Q$=diag([10 10 10]), $S$=0, $R$=10, $N_D$=80, 0.004% unst. | | | | | | | | | |
| Mean | 11.84 | 3.79 | 3.50 | 13.22 | 6.71 | 5.04 | 0.18 | 0.17 | 0.16 |
| Var | 18.46 | 2.91 | 1.92 | 42.56 | 11.60 | 5.61 | | | |
| $\hat{n}$=4, $Q$=diag([10 10 10]), $S$=0, $R$=10, $N_D$=80, 31% unst. | | | | | | | | | |
| Mean | 9.36 | 3.58 | 3.43 | 11.78 | 6.01 | 4.73 | 0.28 | 0.16 | 0.13 |
| Var | 20.05 | 1.43 | 0.82 | 36.84 | 6.49 | 1.51 | | | |
| $\hat{n}$=5, $Q$=diag([10 10 10]), $S$=0, $R$=10, $N_D$=80, 71% unst. | | | | | | | | | |
| Mean | 8.75 | 4.43 | 4.04 | 11.31 | 7.13 | 5.29 | 3.51 | 1.18 | 0.13 |
| Var | 17.53 | 4.23 | 1.96 | 35.8 | 14.5 | 3.27 | | | |

**Table 1:** Monte Carlo simulations for a 3rd order system: see text for details.

| | IR: $k = 10$ | | | IR: $k = 50$ | | | mean(var($\tilde{A}_{ij}$)) | | |
|---|---|---|---|---|---|---|---|---|---|
| | SIr | CM | Reg | SIr | CM | Reg | SI | CM | Reg |
| $\hat{n}$=10, $Q$=diag([1 1 1]), $S$=0, $R$=1, $N_D$ = 80, 72% unst. | | | | | | | | | |
| Mean | 3.34 | 2.01 | 2.04 | 4.66 | 3.71 | 3.25 | 0.14 | 0.12 | 0.07 |
| Var | 3.19 | 0.40 | 0.37 | 4.59 | 3.42 | 1.11 | | | |
| $\hat{n}$=10, $Q$=diag([10 10 10]), $S$=0, $R$=10, $N_D$ = 80, 80% unst. | | | | | | | | | |
| Mean | 4.84 | 5.20 | 4.23 | 6.48 | 7.99 | 5.79 | 0.13 | 0.11 | 0.07 |
| Var | 1.81 | 3.05 | 1.36 | 2.61 | 4.39 | 2.51 | | | |
| $\hat{n}$=12, $Q$=diag([1 1 1]), $S$=0, $R$=1, $N_D$ = 80, 99% unst. | | | | | | | | | |
| Mean | 3.17 | 2.65 | 2.59 | 4.47 | 4.66 | 4.08 | 8.41 | 3.19 | 0.09 |
| Var | 3.38 | 2.56 | 1.13 | 4.78 | 4.30 | 3.20 | | | |
| $\hat{n}$=10, $Q$=diag([1 1 1]), $S$=0, $R$=1, $N_D$ = 160, 58% unst. | | | | | | | | | |
| Mean | 2.98 | 1.15 | 1.18 | 4.20 | 2.18 | 1.99 | 0.11 | 0.10 | 0.07 |
| Var | 3.01 | 0.10 | 0.10 | 4.55 | 0.48 | 0.47 | | | |
| $\hat{n}$=10, $Q$=diag([10 10 10]), $S$=0, $R$=10, $N_D$ = 160, 78% unst. | | | | | | | | | |
| Mean | 4.20 | 2.84 | 2.72 | 5.63 | 4.77 | 3.90 | 0.16 | 0.13 | 0.07 |
| Var | 2.17 | 0.47 | 0.50 | 3.52 | 2.51 | 1.58 | | | |
| $\hat{n}$=12, $Q$=diag([1 1 1]), $S$=0, $R$=1, $N_D$ = 160, 95% unst. | | | | | | | | | |
| Mean | 1.06 | 1.36 | 1.19 | 1.75 | 2.62 | 2.39 | 1.34 | 0.98 | 0.07 |
| Var | 0.10 | 0.37 | 0.25 | 0.32 | 1.17 | 1.56 | | | |

**Table 2:** Monte Carlo simulations for a 10th order system: see text for details.

| $i$ | $c_i$ | $\rho(\tilde{A}_{c_i})$ | $i$ | $c_i$ | $\rho(\tilde{A}_{c_i})$ |
|---|---|---|---|---|---|
| 0 | 12.50355 | 0.454573 | 16 | 0.249664 | 0.899839 |
| 1 | 5.778337 | 0.624497 | 17 | 0.215949 | 0.902046 |
| 2 | 3.662277 | 0.703507 | 18 | 0.185945 | 0.903517 |
| 3 | 2.607907 | 0.745583 | 19 | 0.158968 | 0.913159 |
| 4 | 1.959669 | 0.760395 | 20 | 0.139211 | 0.936727 |
| 5 | 1.490683 | 0.784821 | 21 | 0.132518 | 0.944995 |
| 6 | 1.159539 | 0.813874 | 22 | 0.130043 | 0.948090 |
| 7 | 0.933313 | 0.834844 | 23 | 0.129105 | 0.949268 |
| 8 | 0.789826 | 0.848591 | 24 | 0.128747 | 0.949719 |
| 9 | 0.679955 | 0.859316 | 25 | 0.128610 | 0.949892 |
| 10 | 0.594060 | 0.867781 | 26 | 0.128557 | 0.949959 |
| 11 | 0.515040 | 0.875583 | 27 | 0.128537 | 0.949984 |
| 12 | 0.444105 | 0.882529 | 28 | 0.128529 | 0.949994 |
| 13 | 0.383372 | 0.888348 | 29 | 0.128526 | 0.949998 |
| 14 | 0.332029 | 0.893072 | 30 | 0.128525 | 0.949999 |
| 15 | 0.287973 | 0.896867 | | | |

**Table 3:** Convergence of $c_i$ and of $\rho(\tilde{A}_{c_i})$ in the iterative algorithm to determine $c_m$ as an alternative for solving the (large) Generalized Eigenvalue problem to determine $c_m$. These simulation results indicate linear convergence.