2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Compressive System Identification in the Linear Time-Invariant Framework

Roland Tóth, Borhan M. Sanandaji, Kameshwar Poolla and Tyrone L. Vincent

***Abstract*— Selection of an efficient model parametrization (model order, delay, etc.) has crucial importance in parametric system identification. It navigates a trade-off between representation capabilities of the model (structural bias) and effects of over-parametrization (variance increase of the estimates). There exists many approaches to this widely studied problem in terms of statistical regularization methods and information criteria. In this paper, an alternative ℓ_1 regularization scheme is proposed for estimation of sparse linear-regression models based on recent results in compressive sensing. It is shown that the proposed scheme provides consistent estimation of sparse models in terms of the so-called oracle property, it is computationally attractive for large-scale over-parameterized models and it is applicable in case of small data sets, i.e., underdetermined estimation problems. The performance of the approach w.r.t. other regularization schemes is demonstrated in an extensive Monte Carlo study.**

***Index Terms*— Compressive Sensing; System Identification; Linear Time-Invariant Systems.**

I. Introduction

A common problem in parametric system identification is to choose an efficient model parameterization, i.e., *model structure*, in terms of model order, delays, parameterized coefficients, etc., which is rich enough to represent the relevant dynamical behavior of the data-generating system, but it contains only a minimal set of unknown parameters to be estimated. The latter is important to achieve minimal variance of the parameter estimates. The underlying trade-off between under- and over-parametrization, i.e., structural bias and variance, has significant impact on the result of the identification cycle and an optimal choice in this trade-off is one of the primary goals of system identification [1].

Order selection and regularization of parametrizations w.r.t. linear regression models is a widely studied subject in identification (see [1], [2]) originating from the classical results in statistics in terms of the *Akaike Information Criterion* (AIC) [3] and the *Bayesian Information Criterion* (BIC) [4]. More recently, statistical regularization (shrinkage) methods have been developed like the *Non-Negative Garrote* (NNG) or the *Least Absolute Shrinkage and Selection Operator* (LASSO) [5]–[7], or the *Ridge Regression* and the *Elastic Net* methods [8]. The NNG was applied in the context of identification of *Linear Time-Invariant* (LTI) *Auto Regressive with eXogenous input* (ARX) models in [9].

The AIC and the BIC approaches can have a significant computational load as they are based on a combinatorial search scheme, while the LASSO and the NNG utilize convex optimization. However, for the latter approaches, a search over a weighting (regularization) parameter may still be required. In addition, theory for predicting the finite data performance of the LASSO and the NNG in the context of ARX models still appears to be underdeveloped.

The trade-off problem of parametrization significantly increases in difficulty when the data-generating system has a sparse representation, e.g. in discrete time it is described by a difference equation with only a few difference terms with nonzero coefficients, or in case of *Multiple-Input-Multiple-Output* (MIMO) models where certain *Input-Output* (IO) directions have much lower order than others. This commonly results in polynomial IO models with only a (relatively) few nonzero terms. In general, large-scale over-parametrization increases computational load and sensitivity for the choice of the regularization parameter in the above presented shrinkage methods [7].

Another problem arises when only a few data points are available compared to the size of the parameterization. This is often the case for slow sampling rates or when the input is exciting over only a limited time interval like in the case of step responses of process systems. Traditional identification and model structure selection has proven to be unreliable in these cases and commonly estimates tend to be biased or have large variance. Most theoretical results on the stochastic properties of the parameter estimates concentrate on the asymptotic case, with only a few results concerning the finite-data set case. It appears that classical LTI identification has severe restrictions w.r.t. these scenarios with little work done on non-well-posed or underdetermined problems.

In this paper, we aim to explore model parameter estimation assisted by parameter regularization with particular emphasis on the case when the number of data points is on the same order as the model structure parameters. In fact, we investigate the usefulness of recent results of the *Compressive Sensing* (CS) field in this context. CS is an emerging framework for optimization/estimation of sparse parameter representations, however only with some preliminary results in system ID. Our objective is to propose an efficient *Compressive System Identification* (CSI) of LTI dynamical systems in the *Prediction Error Minimization* (PEM) framework and to demonstrate its usefulness for:

1) Optimal selection of polynomial IO model structures.

R. Tóth is with the Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, The Netherlands, email: r.toth@tudelft.nl. Supported by NWO (grant no. 680-50-0927).

K. Poolla is with the Berkeley Center for Control and Identification, University of California at Berkeley, 5105 Etcheverry Hall, Berkeley, CA 94720-1740, USA, email: poolla@berkeley.edu. Supported by NSF (grant no. ECCS-0925337) and OOF991-KAUST US LIMITED (award no. 025478).

B. M. Sanandaji and T. L. Vincent are with the Department of Electrical Engineering and Computer Sciences, Colorado School of Mines, Golden, CO 80401, USA email: bmolazem@mines.edu; tvincent@mines.edu. Supported by NSF (grant no. CNS-0931748).

2) Delivering optimal parameter estimates in case of underdetermined identification scenarios.

The paper is organized as follows: In Sec. II, basic results of CS are introduced and an ℓ_1 sparse (CSI) estimator is derived which is studied in the context of PEM identification of linear regression models in Sec. III. Next, the consistency properties of the introduced approach are analyzed. Connection of the CSI method with the LASSO and the NNG sparse estimators is explored in Sec. IV and fruitful insights are established. In Sec. V, performance of the CSI is compared to other regularization schemes in a Monte Carlo study.

II. Compressive sensing

As a first step, the core problem setting of CS is introduced from the perspective of linear regression models and the main results of this framework, which will be used in the rest of the paper, are presented.

Consider the discrete-time signal $y : \mathbb{Z} \to \mathbb{R}^m$ given as

$$y(k) = \sum_{i=1}^{n} \theta_{\mathrm{o},i}\psi_i(k), \tag{1}$$

where $\{\psi_i(k)\}_{i=1}^n$ is a set of normalized (orthogonal) basis functions in an appropriate dot product space over $(\mathbb{R}^m)^{\mathbb{Z}}$, with inner product $\langle\cdot,\cdot\rangle$, and the constant expansion coefficients $\theta_{\mathrm{o}} = [\ \theta_{\mathrm{o},1} \quad \dots \quad \theta_{\mathrm{o},n}\ ]^\top \in \mathbb{R}^n$ satisfying $\theta_{\mathrm{o},i} = \langle y(k), \psi_i(k)\rangle$. Note that this is a classical definition of linear regression models where each $\theta_{\mathrm{o},i}$ corresponds to a parameter to be estimated while $\psi_i(k)$ are the regressor terms that can, but not restricted to, contain lagged versions of the IO signals of the model.

Assume that $\{\psi_i(k)\}_{i=1}^n$ is an over-complete basis set w.r.t. $y(k)$, yielding that θ_{o} is sparse. A vector $x \in \mathbb{R}^n$ is called sparse if $\|x\|_{\ell_0} \ll n$ where $\|\,.\,\|_{\ell_0}$ returns the number of nonzero elements of x. For a given $\mathcal{T} \subset \mathbb{I}_n \triangleq \{1,\dots,n\}$ with τ elements, let $x_{\mathcal{T}}$ denote the τ-sparse projection of x, where $[x_{\mathcal{T}}]_i = [x]_i$ if $i \in \mathcal{T}$ and 0 otherwise. We can relax the previous definition of sparsity by calling x to be compressible if $\exists \mathcal{T} \subset \mathbb{I}_n$ with $\mathrm{Card}(\mathcal{T}) = \tau$ s.t. $\|x - x_{\mathcal{T}}\|_{\ell_1} \approx 0$.

Assume that $y(k)$ is available for a time interval $1 \le k \le N$ and define $Y = [\ y(1)^\top \quad \dots \quad y(N)^\top\ ]^\top$ with

$$\Psi \triangleq \begin{bmatrix} \psi_1(1) & \dots & \psi_n(1) \\ \vdots & \ddots & \vdots \\ \psi_1(N) & \dots & \psi_n(N) \end{bmatrix} \triangleq \begin{bmatrix} \varphi_1^\top \\ \vdots \\ \varphi_N^\top \end{bmatrix}. \tag{2}$$

Note that $Y = \Psi\theta_{\mathrm{o}}$. The basic objective in CS is to represent the signal $y(k)$ by computing a θ with maximal sparsity. This corresponds to minimizing the ℓ_0 pseudo-norm of θ under the constraint that $Y = \Psi\theta$. As this so-called *sparse optimization problem* is non-convex and NP hard, a fruitful alternative is a convex relaxation based on the ℓ_1 norm [10]:

$$\underset{\theta\in\mathbb{R}^n}{\text{minimize}} \quad \|\theta\|_{\ell_1}, \tag{3a}$$

$$\text{s.t.} \quad y(k) = \varphi_k^\top\theta,\ \forall k \in \{1,\dots,N\}. \tag{3b}$$

Note that (3a-b) is a classical linear programing problem which is efficiently solvable, even in case of $n \gg N$, by greedy algorithms like *Matching Pursuit* (MP) [11] or by standard optimization techniques [12] using interior-point methods based solvers like SeDuMi [13] or more efficient algorithms tuned to this problem: *l1_ls* [14] or ℓ_1*-magic* [15].

Now consider that measurement of y is effected by noise, i.e., $\tilde{y}(k) = y(k) + e(k)$, where $e(k)$ is an arbitrary bounded noise process, e.g. white and $e(k) \in \mathcal{N}(0, I_{m\times m}\sigma_{\mathrm{e}}^2)$ for all $k \in \mathbb{I}_N$. This corresponds to $Y = \Psi\theta_{\mathrm{o}} + E$ where $E = [\ e(1)^\top \quad \dots \quad e(N)^\top\ ]^\top$. Then (3a-b) is modified as

$$\underset{\theta\in\mathbb{R}^n}{\text{minimize}} \quad \|\theta\|_{\ell_1}, \tag{4a}$$

$$\text{s.t.} \quad \|\tilde{y}(k) - \varphi_k^\top\theta\|_{\ell_2} < \varepsilon,\ \forall k \in \{1,\dots,N\}. \tag{4b}$$

where $\varepsilon > 0$ is a priori chosen. This estimator, what we will call the CSI method, is again a convex problem (a second-order cone program) and can be solved efficiently [16], e.g. by the above mentioned solvers or MP. Algorithms also exist to optimize the value of ε and avoid cases of noise over-fitting or under-fitting [17].

By establishing conditions on Ψ, like the *Restricted Isometry Property* (RIP) [1], reconstruction of a compressible (τ-sparse) θ_{o} can be guaranteed [19]. However, if the regression matrix Ψ contains columns that are the inputs and outputs of a dynamical system, then the RIP conditions are much more difficult to verify (due to the correlation between input and output measurements). Thus, we will consider an alternative result in CS to address the recovery condition. We will see that this is essential to prove consistency of the sparse estimator (4a-b) in an identification setting.

Theorem 1 (Recovery, [20], [21]): Consider $Y \in \mathbb{R}^N$ generated by $Y = \Psi\theta_{\mathrm{o}} + E$ with $\Psi \in \mathbb{R}^{N\times n}$, $\theta_{\mathrm{o}} \in \mathbb{R}^n$ and $E \in \mathbb{R}^N$ being a stochastic noise sequence with $\|E\|_{\ell_2} = \varepsilon_{\mathrm{o}}$ bounded. Assume w.l.o.g. that each column of Ψ: ψ_i, $i \in \{1,\dots,n\}$ is nonzero and Ψ is normalized in the sense that $\psi_i^\top\psi_i = 1$. Let $0 < \|\theta_{\mathrm{o}}\|_{\ell_0} = \tau < n$ with $\mathrm{Sup}(\theta_{\mathrm{o}}) = \mathcal{T}$ and denote $\Psi_{\mathcal{T}}$ the matrix formed from the columns of Ψ listed in $\mathcal{T}$. Let $\Psi_{\mathcal{T}}^+ = (\Psi_{\mathcal{T}}^\top\Psi_{\mathcal{T}})^{-1}\Psi_{\mathcal{T}}^\top$ be the Moore-Penrose pseudo-inverse of $\Psi_{\mathcal{T}}$ and $\|\,.\,\|_{p,q}$ the p,q-matrix-operator norm. Sufficient conditions that the solution $\hat{\theta}$ of (4a-b) obeys

$$\|\theta_{\mathrm{o}} - \hat{\theta}\|_{\ell_2} \le \varepsilon\|\Psi_{\mathcal{T}}^+\|_{2,2}, \tag{5a}$$

$$\mathrm{Sup}(\hat{\theta}) \subseteq \mathcal{T}, \tag{5b}$$

are

$$\mathrm{ERC}(\Psi,\mathcal{T}) \triangleq 1 - \max_{i\in\mathbb{I}_n\setminus\mathcal{T}} \|\Psi_{\mathcal{T}}^+\psi_i\|_{\ell_1} > 0, \tag{6a}$$

$$\varepsilon \ge \varepsilon_{\mathrm{o}}\sqrt{1 + \left(\frac{\|\Psi_{\mathcal{T}}^+\|_{2,1}}{\mathrm{ERC}(\Psi,\mathcal{T})} \cdot \max_{i\in\mathbb{I}_n} \frac{|E^\top\psi_i|}{\varepsilon_{\mathrm{o}}}\right)^2}. \tag{6b}$$

Condition (6a) can be interpreted as the largest absolute value of the cosine between different columns of Ψ while (6b) is based on the correlation between the noise e and ψ_i. As in practice the sparsity-structure of θ_{o} is unknown, computable bounds of (6a) in terms of *cumulative coherence* are commonly applied [21]. In conclusion, the estimation scheme (4a-b) has the following advantages:

- Applicable even in case of serious over-parametrization.
- Computationally attractive.
- Recovery of θ_{o} in the sense of Th. 1 is guaranteed.

[1] If Ψ contains iid sub-Gaussian elements and $N \sim \|\theta_{\mathrm{o}}\|_{\ell_0}\log(n)$, then Ψ satisfies the RIP condition with high probability [18].

III. Compressive system identification

In this section the estimation problem (4a-b) is formulated in the classical LTI *prediction-error* setting and consistency conditions are established. Due to space restrictions, the discussion is restricted to polynomial ARX models in the *Single-Input Single-Output* (SISO) case, however most of the results generalizes to the MIMO case.

A. Prediction error approach: the ARX case

In PEM approaches of LTI system identification, the data-generating system $\mathcal{M}_{\rm o}$ and the model structure is often considered in a polynomial IO representation [1]. In the ARX case, $\mathcal{M}_{\rm o}$ is defined as

$$A_{\rm o}(q^{-1})y(k) = B_{\rm o}(q^{-1})u(k) + e_{\rm o}(k), \tag{7}$$

where $u, y : \mathbb{Z} \to \mathbb{R}$ are the input and output signals respectively, q^{-1} is the *backward time-shift* operator, i.e., $q^{-1}u(k) = u(k-1)$, $e_{\rm o}(k)$ is a zero mean white noise process, $A_{\rm o}$, $B_{\rm o}$ are polynomials with $n_{\rm a} = \deg(A_{\rm o}), n_{\rm b} = \deg(B_{\rm o}) \geq 0$ and for all roots λ of $\xi^{n_{\rm a}}A_{\rm o}(\xi^{-1})$, $|\lambda| < 1$ (stable noise model).

To capture/approximate (7) based on a measured data sequence $\mathcal{D}_N = \{y(k), u(k)\}_{k=1}^N$ with $N > 0$, the model structure is also defined in the form of (7) with the same conditions and parameterized polynomials:

$$A(q^{-1},\theta) = 1 + \sum_{i=1}^{n_{\rm a}} a_i q^{-i},\ B(q^{-1},\theta) = \sum_{j=0}^{n_{\rm b}} b_j q^{-j}, \tag{8}$$

with parameters

$$\theta = [\ a_1 \quad \dots \quad a_{n_{\rm a}} \quad b_0 \quad \dots \quad b_{n_{\rm b}}\ ] \in \mathbb{R}^{n_{\rm a}+n_{\rm b}+1}.$$

The parameterized model in this form is given as $\mathcal{M}_\theta : (A(q^{-1},\theta), B(q^{-1},\theta))$. It is possible to show (see [1]) that w.r.t. (7), the conditional expectation of $y(k)$ in the ℓ_2 sense under the information set of $\mathcal{D}_{k-1} \cup \{u(k)\}$ is equal to

$$y(k|k-1) = (1 - A_{\rm o}(q^{-1}))y(k) + B_{\rm o}(q^{-1})u(k). \tag{9}$$

The basic philosophy of PEM based identification is that w.r.t. a given model set $\mathcal{M} = \{\mathcal{M}_\theta \mid \theta \in \mathbb{R}^{n_\theta}\}$ and a data set $\mathcal{D}_N$, find θ such that the one-step-ahead predictor

$$\hat{y}_\theta(k|k-1) = (1 - A(q^{-1},\theta))y(k) + B(q^{-1},\theta)u(k), \tag{10}$$

associated with $\mathcal{M}_\theta$ provides a *prediction error*

$$e_\theta(k) = y(k) - \hat{y}_\theta(k|k-1), \tag{11}$$

which resembles a zero mean white noise "as much as possible". In this sense, the estimation problem of θ w.r.t. a given $\mathcal{D}_N$ is commonly formulated in terms of the minimization of the *mean-squared prediction error* criterion:

$$W(\theta, \mathcal{D}_N) = \frac{1}{N}\sum_{k=0}^{N} e_\theta^2(k) = \frac{1}{N}\|e_\theta\|_{\ell_2}^2 \tag{12}$$

resulting in the *least-squares* (LS) parameter estimate:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^{n_\theta}} W(\theta, \mathcal{D}_N). \tag{13}$$

By introducing the regressor

$$\varphi_k^\top = [\ -y(k-1) \quad \dots \quad -y(k-n_{\rm a}) \quad u(k) \quad \dots \quad u(k-n_{\rm b})\ ], \tag{14}$$

(10) corresponds to a linear regression

$$\hat{y}_\theta(k|k-1) = \varphi_k^\top \theta. \tag{15}$$

Under the ℓ_2 cost function (12), (13) is a quadratic-optimization problem with an analytical solution.

It is a well known that the ℓ_2 optimization (13) in case of over-parametrization, i.e., sparsity of $\theta_{\rm o}$, distributes power of $\hat{\theta}$ to superfluous parameters (governed by the Tykhonov regularization theory). This means that $\hat{\theta}$ in (13) is almost never sparse [22]. Sensitivity for this property scales with the power of $e_{\rm o}$ and $1/N$. This phenomenon is responsible for an increased variance of the estimate. Thus in these situations, minimization of the support of θ, i.e., the $\|\theta\|_{\ell_0}$ norm becomes important.

B. Compressive identification for ARX models

Next, we investigate how the possible sparsity of $\theta_{\rm o}$ can be exploited/enforced during the optimization (13) and in this way increase the accuracy of the estimate. Here we investigate only the convex optimization based solution of (4a-b) in this context. A block orthogonal-MP based solution is explored in [23].

To formulate this estimation problem in the CS setting, introduce $Y = [\ y^\top(1) \quad \dots \quad y^\top(N)\ ]^\top$ and $\Psi = [\ \varphi_1^\top \quad \dots \quad \varphi_N^\top\ ]$ based on a given $\mathcal{D}_N$ and denote the columns of Ψ with ψ_i. Based on (15), we can write

$$Y = \Psi\theta + E_\theta, \tag{16}$$

where $E_\theta = [\ e_\theta(1) \quad \dots \quad e_\theta(N)\ ]^\top$ is the prediction error. According to the PEM philosophy, our aim is to find the best sparse θ such that E_θ is a sequence of iid samples from a zero mean distribution. Assume that this distribution is Gaussian, i.e., $e_{\rm o}(k) \in \mathcal{N}(0, \sigma_{\rm e}^2)$ yielding that $\|E_\theta\|_{\ell_2} \approx \sqrt{N}\sigma_{\rm e}$. Based on the CS setting, estimation of a sparse $\theta_{\rm o}$ w.r.t. (16) can be efficiently achieved via the convex problem (4a-b).

To establish results about consistency of the proposed sparse estimator, assume that for a given data record $\mathcal{D}_N$, the true underlying noise bound $\varepsilon_{\rm o} = \|E_{\theta_{\rm o}}\|_{\ell_2} > 0$ is known. Then the following theorem holds:

Theorem 2 (ARX recovery): Let $\mathcal{D}_N$ be generated by an ARX model (7) with a sparse parameter vector $\theta_{\rm o} \in \mathbb{R}^{n_\theta}$ with support $\mathcal{T}$. Let $\mathcal{M}_{\rm o} \in \mathcal{M}$ and Ψ be constructed from $\mathcal{D}_N$ according to (14). Assume that in $\mathcal{D}_N$, u is white with $u(k) \in \mathcal{N}(0, \sigma_{\rm u}^2)$, $\sigma_{\rm u} > 0$ giving that $\Psi_\mathcal{T}^\top\Psi_\mathcal{T} \succ 0$ with probability 1, then the expected value of the normalized form of Ψ satisfies (6a) for large enough N and hence in terms of Th. 1, with $\varepsilon \geq \varepsilon_{\rm o}$ chosen according to (6b), recovery of $\theta_{\rm o}$ holds.

For a proof see the Appendix. This concludes that it is possible to design u, i.e., $\mathcal{D}_N$, such that the recovery condition of Th. 1 is satisfied. In particular, the proof of Th. 2 reveals that a necessary condition for N is

$$\max_{i \in \mathbb{I}_n \setminus \mathcal{T}} \|Q_i\|_{\ell_1}^2 + \frac{2}{\pi}\left(\operatorname{trace}\left(P_i^{1/2}\right)\right)^2 < N, \tag{17}$$

where Q_i and P_i are defined by (36a-b), which, if $\theta_{\rm o}$ is sparse, can be significantly smaller than n_θ. Note that the whiteness and Gaussian distribution of u is a technical necessity, but it is expected that these conditions can be further relaxed.

IV. Comparison to other sparse estimators

Next we investigate how the introduced sparse estimation scheme compares to other sparse estimators or model structure selection approaches in the PEM framework.

A. AIC & BIC criterion

The AIC criterion is defined as

$$\mathrm{AIC} = 2\frac{n_\theta}{N} + \log\Big(\frac{\|e_\theta\|_{\ell_2}^2}{N}\Big), \tag{18}$$

while the BIC criterion (in case of ARX, where e_o has a normal distribution) is:

$$\mathrm{BIC} = \frac{n_\theta \log(N)}{N} + \log\Big(\frac{\|e_\theta\|_{\ell_2}^2}{N}\Big). \tag{19}$$

In practice, these criteria are evaluated for LS parameter estimates generated by all possible selection of at most n_θ columns of the regressor matrix Ψ. This correspond to exploring all possible sparse solutions by solving $\sum_{k=1}^{n_\theta}\binom{n_\theta}{k} = 2^{n_\theta}-1$ linear regression problems. Then by evaluating the BIC or AIC on the estimation or validation data w.r.t. each estimate, the most likely model structure of the system follows at the minimum of these criteria. This means that the computational load exponentially grows with the number of regression terms (NP-hard problem). In practice, these types of methods are implemented in a stepwise fashion, through forward selection or backward elimination. Because of the myopic nature of the stepwise algorithm, these implementations are known to be suboptimal [24].

B. Sparse estimators: LASSO and NNG

To avoid the computational explosion and to provide a compact estimator, it is attractive to combine minimization of (12) with the minimization of the support of θ, i.e., $\|\theta\|_{\ell_0}$. However, (12) and $\|\theta\|_{\ell_0}$ can not be minimized simultaneously (same target variable) and the ℓ_0 problem is NP hard. Thus using the same motivation as in CS, a convex relaxation can be introduced in terms of

$$\text{minimize} \quad \|\theta\|_{\ell_1}, \tag{20}$$

which can still guarantee sparsity in a computationally more attractive setting. The idea is to combine the optimization problems (20) and (13) by using (20) as a constraint: $\|\theta\|_{\ell_1} < \varepsilon$, where ε is given, or by using the weighted sum of (12) and $\|\theta\|_{\ell_1}$ resulting in a set of regressor regularization/shrinkage methods like the NNG and the LASSO.

The LASSO method w.r.t. a linear regression model (16) is formulated as

$$\underset{\theta\in\mathbb{R}^n}{\text{minimize}} \quad \|\tilde{y}(k) - \varphi_k^\top\theta\|_{\ell_2},\ \forall k \in \{1,\dots,N\}, \tag{21a}$$

$$\text{s.t.} \quad \|\theta\|_{\ell_1} \le \varepsilon, \tag{21b}$$

corresponding to a quadratic programing problem where ε is usually obtained by lowering it iteratively based on cross-validation. In practice, this approach is usually implemented by using greedy algorithms [6], but more advanced piecewise-linear solution path based methods also exist [25].

The NNG approach, instead of affecting the estimation of θ directly, penalizes the LS solution by attaching weights to it, which in turn are regularized. Thus, given the least-squares estimate $\hat{\theta}_N$ of (16), the NNG problem is formulated as

$$\min_{w} \ \sum_{k=1}^{N}\Big(y(k) - \sum_{i=1}^{n_\theta} w_i\psi_i(k)\hat{\theta}_i\Big)^2 + \lambda\sum_{i=1}^{n_\theta} w_i, \tag{22a}$$

$$\text{s.t. } w \succeq 0, \tag{22b}$$

where λ is the model complexity parameter, $\psi_i(k)$ is the i-th element of φ_k and $w \triangleq [\ w_1 \ \ \dots \ \ w_{n_\theta}\ ]^\top$ are the weights. For a given λ, (22a-b) is a convex optimization problem in w, and the delivered estimate is $\tilde{\theta} = w \odot \theta$ with $\odot$ being the *Hadamard product*. As λ increases, the weights of the less important regressors will shrink, and finally end up exactly at zero. This results in less and less complex model estimates, as long as the overall fit of the estimate on the available (validation) data is still acceptable. The fit itself can be calculated in terms of any error measure or the BIC or AIC criterion. An efficient way to implement this strategy is to use a path following parametric estimation, which calculates a piecewise affine solution path for λ [9]. The NNG is reported to be more effective in recovering the sparsity structure of θ_o than the LASSO. However a particular drawback of this approach is that it can not be applied when $N < n_\theta$. To overcome this drawback, in [7], the *Ridge regression* is suggested to be used as an initial estimate.

These sparse estimators have the following property:

Property 1 (Oracle): If $N \to \infty$ and the data-record is persistently exciting w.r.t. the considered $\mathcal{M}$, where $\mathcal{M}_\mathrm{o} \in \mathcal{M}$ corresponding to θ_o with support $\mathcal{T}$, then the parameter estimate $\hat{\theta}$ satisfies that the probability $P(r(\theta_\mathrm{o}) = r(\hat{\theta})) = 1$, where $[r(\theta)]_i = 1$ if $\theta_i \neq 0$ while $[r(\theta)]_i = 0$ if $\theta_i = 0$, and $\hat{\theta}_i = \theta_{\mathrm{o},i} + \mathcal{O}(\sigma_\mathrm{e})$ for $i \in \mathcal{T}$.

The oracle property implies that asymptotically, the correct support is estimated with probability one. This would appear to be a very desirable property, yet the same property also implies that the worst-case asymptotic squared error decreases more slowly than of the LS solution:

Property 2 ([26]): Suppose a sparse estimator fulfills the oracle property. If $N \to \infty$, then $P(r(\theta_\mathrm{o}) = r(\hat{\theta})) = 1$, however the maximal risk associated with the identification criterion diverges

$$\sup_{\theta_\mathrm{o}\in\mathbb{R}^{n_\theta}} \mathbb{E}\{N(\hat{\theta}-\theta_\mathrm{o})^\top(\hat{\theta}-\theta_\mathrm{o})\} \to \infty, \tag{23}$$

while in case of the LS solution

$$\sup_{\theta_\mathrm{o}\in\mathbb{R}^{n_\theta}} \mathbb{E}\{N(\hat{\theta}-\theta_\mathrm{o})^\top(\hat{\theta}-\theta_\mathrm{o})\} \to \mathrm{Trace}(Q^{-1}), \tag{24}$$

where $Q = \frac{1}{N}\sum_{k=1}^{N}\varphi_k^\top\varphi_k$.

Although these results are asymptotic, and thus cannot be truly translated to the finite data case, they suggest that the performance of sparse estimators is not uniform w.r.t. θ_o, and that an inherent bias always exists when the oracle property holds. Nevertheless, in the finite data case there can be significant advantages to use sparsity enhancing estimators.

C. Comparison to the CSI approach

AIC and BIC have significant computational load compared to the convex minimization problem of (4a-b) for large n_θ. However, AIC and BIC are expected to deliver more accurate selection as all possible combinations and hence all

possible sparse LS solutions for the estimation of θ are tested. In this respect (4a-b) presents a computationally attractive solution just like the LASSO and the NNG.

In comparison with the LASSO approach, (4a-b) corresponds to an alternative solution for the same sparse estimation problem. Note that by solving (4a-b), the identification criterion is

$$W(\theta, \lambda, \mathcal{D}_N) = \|\theta\|_{\ell_1} + \lambda(\frac{1}{N}\|e_\theta\|_{\ell_2} - \varepsilon), \qquad (25)$$

with both θ and $\lambda \geq 0$ as optimization variables. This provides a *sum-of-norms* type of criterion function where λ, i.e., the regularization parameter is optimized. Contrary to other type of sparse estimators, the "optimal" regularization parameter is directly delivered in this case via the choice of ε, giving a straightforward interpretation of λ in terms of the user chosen error bound. In terms of objectives, while in (4a-b), the ℓ_1 norm of the estimated parameter vector is minimized to achieve the best sparsity level for a predescribed prediction error controlled via ε, in the LASSO case, the ℓ_2 cost of the prediction error is minimized for a given sparsity level. As the ℓ_1 norm of the optimal estimate for θ is unknown, it is hard in practice to guess a good estimate for ε in the LASSO case, while in the CSI case we know that the expected error is white and its variance is much easier to estimate. This means that it is practically more attractive to use (4a-b) as it is generally expected to be easier to achieve recovery of the unknown sparse structure of $\theta_{\rm o}$. However, if ε is optimally chosen, then the two optimization problems are equivalent.

Comparison to the NNG shows that the re-weighting approach is somewhere in between the LASSO provided optimization problem and (4a-b). However, particular disadvantages of the NNG is its sensitivity for the undetermined regression case and the non-trivial relationship between the expected prediction error, sparsity level of the estimate and the value of λ. Thus the solution needs to be explored for all λ which is done via a sub-optimal piecewise solution path. Depending on the size of the regression problem, the *Signal to Noise Ratio* (SNR) and the sparsity level of $\theta_{\rm o}$, this can result in varying computational time ranging form a few seconds to hours. The CSI is empirically observed to more efficiently recover the sparsity structure and it is also applicable in the underdetermined case (see Sec. V).

Finally, to show that the CSI satisfies the oracle property, like the NNG and the LASSO, consider consistency in terms of Th. 1 when $\sigma_{\rm e} \to 0$ and ε is chosen as the minimal value satisfying (6a). $\sigma_{\rm e} \to 0$ implies that $\varepsilon_{\rm o} = \|E_{\theta_{\rm o}}\|_{\ell_2} \to 0$. As a consequence, $P(\hat{\theta} = \theta_{\rm o}) \to 1$ as $\sigma_{\rm e} \to 0$ implies that $\varepsilon \to 0$. Now let $\sigma_{\rm e} > 0$ and consider $N \to \infty$. Then in terms of the proof of Th. 2, $\mathrm{ERC}(\Psi, \mathcal{T}) \to 1$ and the minimum of ε converges to $\varepsilon_{\rm o}$. This implies that $\mathrm{Sup}(\hat{\theta}) = \mathcal{T}$ which proves that the proposed sparse estimator satisfies the *oracle property*. On the other hand, if $N \to \infty$ then in (6b) $\|E_{\theta_{\rm o}}\|_{\ell_2} = \varepsilon_{\rm o} \to \infty$ leaving (5a) unbounded. This yields that even if $P(\hat{\theta}_i = 0) = 1$ for $i \notin \mathcal{T}$, at the same time $P(\hat{\theta} = \theta_{\rm o}) = 0$. This points out that sparse recovery has got a price for CSI as well in terms of maximizing the loss for $N \to \infty$ (see Property 2). To decrease the effect of this property, the following strategy can be used:

1) Wr.t. a given ARX model structure $\mathcal{M}_\theta$ and data set $\mathcal{D}_N$, estimate $\hat{\theta}$ according to (4a-b) where the regressor matrix Ψ is generated according to (14).
2) Based on a threshold $0 < \varepsilon_* \ll 1$ select a subset $\mathcal{T}$ of the support of $\hat{\theta}$ such that $\|\hat{\theta}_{\mathcal{T}} - \hat{\theta}\|_{\ell_1} < \varepsilon_* \|\hat{\theta}\|_{\ell_1}$.
3) Estimate $\tilde{\theta}$ based on a LS estimate with $\Psi_{\mathcal{T}}$.

This means that the oracle property of (4a-b) is exploited to select the correct columns of Ψ. As recovery of the underlying support of $\theta_{\rm o}$ holds with an overwhelming probability under minor conditions on $\mathcal{D}_N$, thus the LS estimate w.r.t. $\Psi_{\mathcal{T}}$ is consistent as $N \to \infty$. However, this holds only for infinite data. For $N < \infty$, there will of course remain the possibility that the estimated support is incorrect, so re-estimation does not fundamentally get around the problem illustrated in Property 2. Yet practical advantages exist, which are explored numerically in the example.

Another remark is that the consistency result has been established based on the optimal choice of ε for (4a-b), i.e., using condition (6b) with $\varepsilon_{\rm o} = \|E_{\theta_{\rm o}}\|_{\ell_2}$, which is not available in practice. Different schemes can be applied to approximate a reasonably good value of ε based on data like an n-section based search starting from an upper bound of $\varepsilon_{\rm o}$ calculated from the estimated noise w.r.t. an LS estimate. For more on the appropriate choice of ε see the recent results in [27].

Finally, it is well known in the LTI literature that ARX models are globally identifiable, also in case of over-parametrization, if $e_{\rm o}$ has a nonzero variance, i.e., $\varepsilon_{\rm o} > 0$ [28]. However in case $e_{\rm o} = 0$, the ARX model structure is not identifiable (locally at $\theta_{\rm o}$) if $\deg(A(q^{-1}, \theta) \neq \deg(A(q^{-1}, \theta_{\rm o})$ due to pole-zero cancelations. This means that consistency requires this assumption if $\sigma_{\rm e} \to 0$.

V. Example

Next the performance of the proposed CSI is compared to the NNG via a representative Monte Carlo study. As the LASSO is considered to be less effective than the NNG and to avoid problems in choosing optimal regularization parameters, comparison is restricted to the CSI and the NNG.

In each simulation, the true system is considered to be a randomly generated stable ARX model with $n_{\rm a} = n_{\rm b} - 1 = 10$, but with $\|\theta_{\rm o}\|_{\ell_0} = 6$ (i.e., 3 nonzero parameters w.r.t. A and 3 w.r.t. B). Furthermore, each model is generated in the sense that the nonzero parameters are in the region $\pm[0.5, 1.5]$ to keep the relative importance of each parameter close to the others. This means that $\theta_{\rm o}$ associated with each of the generated systems is rather sparse and over-parameterization is likely to happen w.r.t. both the model order and input delay. Using randomly generated systems, a Monte Carlo simulation of 100 runs is executed for increasing length of data records $\mathcal{D}_N$ generated by these systems with $N \in [10, 80]$ for a white noise u with $u(k) \in \mathcal{N}(0, 1)$ and $\sigma_{\rm e}^2 = 0.1$ corresponding to an SNR of 55dB (other noise scenarios are not presented due to space restrictions). This means that in each of the 81×100 runs, a new randomly

		MSE		BFR		ℓ_1 error	
N	Method	mean	std	mean	std	mean	std
35	LS-oracle	$1.29 \cdot 10^{-2}$	$4.61 \cdot 10^{-3}$	97.76	1.97	$7.81 \cdot 10^{-2}$	$4.21 \cdot 10^{-2}$
	LS-full	$2.48 \cdot 10^{-2}$	$1.52 \cdot 10^{-2}$	96.27	3.16	2.97	2.23
	CSI-I	$3.19 \cdot 10^{-2}$	$2.58 \cdot 10^{-2}$	96.05	2.64	1.43	1.85
	CSI-I-opt	$1.91 \cdot 10^{-2}$	$1.12 \cdot 10^{-2}$	96.67	2.45	1.02	1.28
	CSI-II-opt	$1.39 \cdot 10^{-2}$	$5.26 \cdot 10^{-3}$	97.64	2.15	$4.54 \cdot 10^{-1}$	1.05
	NNG-BIC	$4.49 \cdot 10^{-2}$	$8.77 \cdot 10^{-2}$	95.06	8.47	$9.11 \cdot 10^{-1}$	1.69
80	LS-oracle	$1.08 \cdot 10^{-2}$	$1.91 \cdot 10^{-3}$	98.61	1.01	$4.25 \cdot 10^{-2}$	$1.97 \cdot 10^{-2}$
	LS-full	$1.32 \cdot 10^{-2}$	$2.55 \cdot 10^{-3}$	97.82	1.24	1.37	$9.58 \cdot 10^{-1}$
	CSI-I	$1.44 \cdot 10^{-2}$	$3.79 \cdot 10^{-3}$	97.26	1.91	$6.79 \cdot 10^{-1}$	1.03
	CSI-I-opt	$1.22 \cdot 10^{-2}$	$2.42 \cdot 10^{-3}$	97.96	1.26	$4.93 \cdot 10^{-1}$	$5.98 \cdot 10^{-1}$
	CSI-II-opt	$1.09 \cdot 10^{-2}$	$2.08 \cdot 10^{-3}$	98.59	1.07	$1.02 \cdot 10^{-1}$	$3.90 \cdot 10^{-1}$
	NNG-BIC	$1.22 \cdot 10^{-2}$	$6.92 \cdot 10^{-3}$	98.20	2.02	$1.14 \cdot 10^{-1}$	$2.78 \cdot 10^{-1}$

TABLE I

MONTE CARLO SIMULATION RESULTS WITH SNR 55DB.

generated system, input and noise realizations are used. The following methods are used to estimate $\theta_{\rm o}$:

- LS-oracle: LS estimate by using the optimal model structure, i.e., $\Psi_{\mathcal{T}}$. This approach is used as a baseline estimate to show the best achievable performance by any regression based estimator in the considered setting.
- LS-full: LS estimate using the full ARX(10,9) model structure (MATLAB toolbox: arx method).
- NNG-BIC: NNG with a piecewise solution path and BIC as a posterior selection of λ using validation data.
- CSI-I: The CSI approach (4a-b), using $\varepsilon = \|E_{\hat{\theta}_{\texttt{LS-full}}}\|_{\ell_2}$, where $\hat{\theta}_{\texttt{LS-full}}$ is obtained by LS-full.
- CSI-I-opt: The CSI approach (4a-b), using an n-section based search for optimizing ε based on validation data. For initialization, $\varepsilon = \|Y\|_{\ell_2}$ is used.
- CSI-II-opt: The CSI-I-opt approach followed by a re-estimation of θ with LS using only the columns of Ψ for which $|[\hat{\theta}_{\texttt{CSI-I-opt}}]_i| \geq \epsilon_* = 0.1$.

Note that LS-oracle can not be applied in practice as the optimal model structure is unknown (part of the identification problem itself). The results are compared in terms of

- The *Mean Squared Error* (MSE) of the prediction:

$$\text{MSE} = \mathbb{E}\{\|y(k) - \hat{y}_{\hat{\theta}}(k\,|\,k-1)\|_{\ell_2}^2\}. \qquad (26)$$

  computed as an average over each 100 runs for a given N and $\sigma_{\rm e}$.
- The *fit score* or the *Best Fit Rate* (BFR) [29]:

$$\text{BFR} = 100\% \cdot \max\left(1 - \frac{\left\|y(k) - \hat{y}_{\hat{\theta}}(k)\right\|_{\ell_2}}{\|y(k) - \bar{y}\|_{\ell_2}}, 0\right), \qquad (27)$$

  where $\bar{y}$ is the mean of y and $\hat{y}_{\hat{\theta}}$ is the simulated model output.
- $\|\hat{\theta} - \theta_{\rm o}\|_{\ell_1}$.

The results w.r.t. the SNR= 55dB case are given in Table I and in Figures 1 and 2. The LS-full is presented for $N \geq n_\theta = 20$ and the NNG-BIC is presented for cases when $N \geq 1.5 n_\theta = 35$ which are built in lower bounds of the used scripts in the identification toolbox. As we can see, the LS-full has a huge bias around $N = 20$ which slowly decreases as N grows, however compared to the LS-oracle, it is still substantial when $N = 80$. On the other hand the NNG-BIC shows worse behavior than the LS-full in terms of MSE for small N, but with a fair BFR and ℓ_1 estimation error. The mean of all error measures rapidly decreases as N-grows converging to the level of the LS-oracle, however as we can see, the standard deviation of these measures even for $N = 80$ is close to the variance of the LS-full estimate which can be recognized as an influence of the initial LS-full estimate.

As we can see the results of the CSI-I are not that impressive compared to the NNG-BIC or to the LS-full, even if it delivers reasonably good estimates for $N < n_\theta$. However, the CSI-I-opt provides results with one magnitude less in all error measures, which clearly indicates that how important is to optimize the error bound ε. This is a general property of any regularization based sparse estimator, i.e., adequate choice of the regularization parameter is crucial to deliver unbiased estimates. However in case of the CSI, it is computationally attractive to optimize ε.

It is also an important observation that re-estimation, i.e., using the sparse estimator only as a model selection tool as in CSI-II-opt, further refines the performance of the estimation scheme. This delivers an estimator which gets the closest to LS-oracle and has smaller bias and variance than the NNG-BIC. In this respect, bias follows by mis-selection of the optimal model structure. As N grows, the gap between these methods decreases in terms of the mean of the error measures, but not in terms of variance. The CSI-II-opt has the advantage of delivering relatively accurate estimates even if the data record is short compared to the parameterization (see $N = 35$ in Table I). In case the model is large scale ($n_\theta > 1000$), this property becomes a serious advantage over other sparse estimation schemes.

The results for other noise cases are not presented here due to space restrictions, but if $\sigma_{\rm e}$ decreases, then the relative performance of CSI-II-opt improves w.r.t. NNG-BIC till identifiability issues starts to play a significant role beyond SNR$>$ 180dB. If $\sigma_{\rm e}$ increases, then the performance of CSI-II-opt and NNG-BIC becomes similar and under SNR$<$ 5dB no significant difference can be observed between them for $N \approx 80$. Note that at SNR$<$ 5dB, recovery of the true sparsity structure $\theta_{\rm o}$ less likely to follow and thus threshold based re-estimation schemes like CSI-II-opt

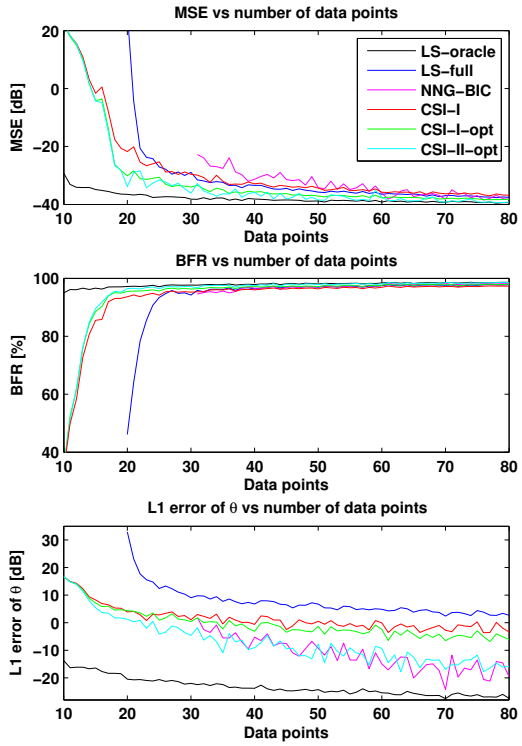

Fig. 1. Monte Carlo simulation results with SNR 55dB.

starts to diverge. If N is further increased, all approaches converge to `LS-oracle` in the means of the error measures. Convergence speed of `CSI-II-opt` and `NNG-BIC` seems to be similar in this study.

VI. Conclusions

Inspired by promising advances in compressive sensing, a new ℓ_1 regularization scheme has been proposed in this paper for the identification of sparse linear-regression models. Recovery and consistency properties of the resulting estimation scheme has been investigated, establishing conditions for finite data sets. Furthermore, it has been shown that the estimator satisfies the oracle property and hence the maximal risk of the estimates is unbounded. To practically overcome this property, a re-estimation scheme has been proposed. Furthermore, the introduced ℓ_1 regularization scheme has been compared to other sparse estimator approaches and it has been shown that its advantages lie in its better accuracy and computational trade-off with a practically sound choice of the regularization parameter.

VII. Acknowledgements

The authors would like to thank H. Hjalmarsson for fruitful discussions on sparse estimators.

VIII. Appendix

Proof: If Ψ is not normalized then $\tilde{\Psi} = \Psi\Sigma^{-1}$ is where

$$\Sigma = I_{n\times n} \cdot [\ \|\psi_1\|_{\ell_2} \quad \dots \quad \|\psi_n\|_{\ell_2}\ ]^\top, \tag{28}$$

and $\tilde{\theta}_\mathrm{o} = \Sigma\theta_\mathrm{o}$ is the corresponding true parameter vector. From this point, assume that Ψ is normalized, i.e., $\|\psi_i\|_{\ell_2} = 1$ for all $i \in \mathbb{I}_n$. Note that $\hat{\eta}_i = \Psi_\mathcal{T}^+\psi_i$ for each $i \in \mathbb{I}_n \setminus \mathcal{T}$ corresponds to the ℓ_2 solution of

$$\psi_i = \Psi_\mathcal{T}\eta_i + V, \tag{29}$$

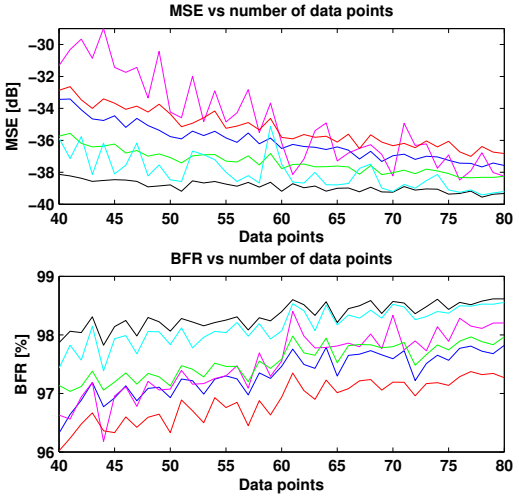

Fig. 2. Zoomed in Monte Carlo simulation results with SNR 55dB.

with $V \in \mathbb{R}^N$. This means that ERC is basically a "measure" of distance between each ψ_i column of Ψ that does not belong to the true support $\mathcal{T}$ and the subspace spanned by the columns of $\Psi_\mathcal{T}$. To have condition (6a) satisfied, $\|\hat{\eta}_i\|_{\ell_1} < 1$ must hold for each $i \in \mathbb{I}_n \setminus \mathcal{T}$.

Let $R(\mathcal{T}) = \Psi_\mathcal{T}^\top\Psi_\mathcal{T}$. The basic condition for (29) to have a unique solution, i.e., $\Psi_\mathcal{T}^+$ to exist, is that $R(\mathcal{T}) \succ 0$, i.e., $\Psi_\mathcal{T}$ is full rank. Assume that the system is well excited in the sense that $R(\mathcal{T}) \succ 0$, which is the classical *persistency of excitation* condition [2] in the ARX case.

If $n_\mathrm{a} = 1$, then $\psi_i = q^{-j}\frac{U}{\|U\|_{\ell_2}}$ with distinct $j \in \mathbb{Z}^+$ for all $i \in \mathbb{I}_n$ where $U \triangleq [\ u(1) \quad \dots \quad u(N)\ ]^\top$. W.l.o.g. assume that u is white and $u(k) \in \mathcal{N}(0, \sigma_\mathrm{u}^2)$ yielding that $\mathbb{E}\{\|U\|_{\ell_2}\} = \sqrt{N}\sigma_\mathrm{u}$. Let $R_{\mathrm{u,u}}(s) \triangleq \mathbb{E}\{u(k)u(k-s)\}$ denote the auto-correlation of u. As $u(k)$ is white, $\eta_{i,\mathrm{o}} = 0$ is the underlying true solution of (29) with $V = \psi_i$ and $R_{\mathrm{u,u}}(s) = \delta(s)\sigma_\mathrm{u}^2/\|U\|_{\ell_2}^2$, where $\delta(\cdot)$ is the Kronecker-delta function. If $N \to \infty$, then based on the central limit theorem it follows (see [1]) that the ℓ_2-solution of (29), i.e., $\hat{\eta}_i = \Psi_\mathcal{T}^+\psi_i$ is consistent and

$$\sqrt{N}\hat{\eta}_i \to \mathcal{N}(0, P_i), \tag{30}$$

with probability 1, where $P_i = N\sigma_\mathrm{u}^2/\mathbb{E}\{\|U\|_{\ell_2}^2\} \cdot C^{-1}$ and C is the correlation matrix of $\Psi_\mathcal{T}$ in this case being equal to $N\sigma_\mathrm{u}^2/\mathbb{E}\{\|U\|_{\ell_2}^2\} \cdot I_{\tau\times\tau}$. This yields that $P_i = I_{\tau\times\tau}$. As a consequence, if $N \to \infty$, then under some minor conditions (see Appendix 9.B in [1]) satisfied by (29) in the considered setting:

$$\mathbb{E}\{\|\hat{\eta}_i\|_{\ell_1}\} = \lim_{N\to\infty} \tau\sqrt{\frac{2}{N\pi}} = 0, \tag{31}$$

based on the fact that $\mathbb{E}\{|x|\} = \sqrt{2/\pi}\sigma$ if $x \in \mathcal{N}(0, \sigma^2)$. This gives the necessary condition for recovery when N is finite, i.e., to have (31) less than 1, that

$$\frac{2}{\pi}\tau^2 < N. \tag{32}$$

Now consider the case when $\Psi_\mathcal{T}^\top$ is formed from the shifted versions of u and y but $\psi_i = q^{-i}U$. Denote

[2]Note that this excitation condition is not for the overall model structure, as $\Psi_\mathcal{T}$ is associated with only the regression vectors of the optimal model structure.

$R_{\mathrm{y,u}}(s) \triangleq \mathbb{E}\{y(k)u(k-s)\}$ the cross-correlation of y w.r.t. u. Note that

$$R_{\mathrm{y,u}}(s) = h_{\mathrm{o}}(s) \star R_{\mathrm{u,u}}(s) = h_{\mathrm{o}}(s)\sigma_{\mathrm{u}}^2 = R_{\mathrm{u,y}}(-s), \quad (33)$$

where $h_{\mathrm{o}}(s)$ is the impulse response of $\mathcal{M}_{\theta_{\mathrm{o}}}$ and $\star$ denotes the discrete-time convolution. Furthermore, denote $R_{\mathrm{y,y}}(s) \triangleq \mathbb{E}\{y(k)y(k-s)\}$ satisfying:

$$R_{\mathrm{y,y}}(s) = \big(h_{\mathrm{o}}(-s) \star h_{\mathrm{o}}(s)\big)\sigma_{\mathrm{u}}^2 + \delta(s)\sigma_{\mathrm{e}}^2. \quad (34)$$

Assume that the columns of $\Psi_{\mathcal{T}}$ are ordered such that

$$\Psi_{\mathcal{T}} = \left[\begin{array}{cccccc} \frac{q^{-\alpha_1}Y}{\|Y\|_{\ell_2}} & \cdots & \frac{q^{-\alpha_{n_{\mathrm{y}}}}Y}{\|Y\|_{\ell_2}} & \frac{q^{-\beta_1}U}{\|U\|_{\ell_2}} & \cdots & \frac{q^{-\beta_{n_{\mathrm{u}}}}U}{\|U\|_{\ell_2}} \end{array} \right].$$

By using the central limit theorem (see [1]):

$$\sqrt{N}\hat{\eta}_i \to \mathcal{N}(Q_i, P_i), \quad (35)$$

with probability 1 if $N \to \infty$, where

$$Q_i = R_*^{-1}(\mathcal{T}) \cdot F_*(\mathcal{T}, i), \quad (36a)$$

$$P_i = (R_*^{-1}(\mathcal{T}))^{\top} F_*^{\top}(\mathcal{T}, i) F_*(\mathcal{T}, i) R_*^{-1}(\mathcal{T}), \quad (36b)$$

$$R_*(\mathcal{T}) \triangleq \lim_{N\to\infty} \mathbb{E}\{R(\mathcal{T})\} = C, \quad (36c)$$

$$F_*(\mathcal{T}, i) \triangleq \lim_{N\to\infty} \mathbb{E}\{\Psi_{\mathcal{T}}^{\top}\psi_i\}, \quad (36d)$$

and C is the correlation matrix of the signals in the columns of $\Psi_{\mathcal{T}}$, i.e., the normalized signals $q^{-i}u$ and $q^{-j}y$, while

$$F_*(\mathcal{T}, i) = \left[\begin{array}{ccc} \frac{NR_{\mathrm{y,u}}(\alpha_1 - i)}{\mathbb{E}\{\|U\|_{\ell_2}\}\mathbb{E}\{\|Y\|_{\ell_2}\}} & \cdots & \frac{NR_{\mathrm{y,u}}(\alpha_{n_{\mathrm{y}}} - i)}{\mathbb{E}\{\|U\|_{\ell_2}\}\mathbb{E}\{\|Y\|_{\ell_2}\}} \\ \frac{NR_{\mathrm{u,u}}(\beta_1 - i)}{\mathbb{E}\{\|U\|_{\ell_2}^2\}} & \cdots & \frac{NR_{\mathrm{u,u}}(\beta_{n_{\mathrm{u}}} - i)}{\mathbb{E}\{\|U\|_{\ell_2}^2\}} \end{array} \right]^{\top}.$$

Note that as $i \notin \mathcal{T}$,

$$F_*(\mathcal{T}, i) = \left[\begin{array}{cccc} \frac{h_{\mathrm{o}}(\alpha_1 - i)}{\|h_{\mathrm{o}}\|_{\ell_2}} & \cdots & \frac{h_{\mathrm{o}}(\alpha_{n_{\mathrm{y}}} - i)}{\|h_{\mathrm{o}}\|_{\ell_2}} & 0 \end{array} \right]^{\top}.$$

Similarly, C is a diagonal dominant positive definite matrix with 1 entries in the diagonal and off-diagonal elements that represent a relative ratio between $h_{\mathrm{o}}(\centerdot)$ and $\|h_{\mathrm{o}}\|_{\ell_2}$. As a consequence, if $N \to \infty$, then

$$\mathbb{E}\{\|\hat{\eta}_i\|_{\ell_1}\} = \frac{1}{\sqrt{N}}\left(\|Q_i\|_{\ell_1} + \sqrt{\frac{2}{\pi}}\mathrm{trace}\left(P_i^{1/2}\right)\right) = 0. \quad (37)$$

Note that in case $\psi = q^{-i}Y$, the same limits hold except $F_*(\mathcal{T}, i)$ is more densely populated and hence $\mathbb{E}\{\|\hat{\eta}_i\|_{\ell_1}\}$ decays to zero slower. Eq. (31) also reveals that a necessary condition for recovery when N is finite, i.e., to have (37) less than 1, is

$$\max_{i \in \mathbb{I}_n \setminus \mathcal{T}} \|Q_i\|_{\ell_1}^2 + \frac{2}{\pi}\left(\mathrm{trace}\left(P_i^{1/2}\right)\right)^2 < N. \quad (38)$$

■

REFERENCES

[1] L. Ljung, *System Identification, theory for the user*, 2nd ed. Prentice-Hall, 1999.

[2] K. Åström and P. Eykhoff, "System identification - a survey," *Automatica*, vol. 7, pp. 123–162, 1971.

[3] H. Akaike, "A new look at the statistical model identification," *IEEE Tran. on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[4] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[5] L. Breiman, "Better subset regression using the nonnegative garotte," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[6] R. Tibshirani, "Regression shrinkage and selection with the Lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[7] M. Yuan and Y. Lin, "On the non-negative garrotte estimator," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 143–161, 2007.

[8] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[9] C. Lyzell, J. Roll, and L. Ljung, "The use of nonnegative garrote for order selection of ARX models," in *Proc. of the 45th IEEE Conf. on Decision and Control*, Cancun, Mexico, Dec. 2008, pp. 1974–1979.

[10] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[11] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Tran. on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[13] J. Sturm, "Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–653, 1999.

[14] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An introduction to compressive sampling," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.

[15] E. J. Candès and J. Romberg, "ℓ_1 magic : Recovery of sparse signals via convex programming," Caltech, Tech. Rep., 2005.

[16] F. Santosa and W. W. Symes, "Linear inversion of band-limited reflection seismograms," *SIAM Journal on Scientific Computing*, vol. 7, no. 4, pp. 1307–1330, 1986.

[17] A. C. Gurbuz, J. H. McClellan, and W. R. Scott, "Compressive sensing for subsurface imaging using ground penetrating radar," *IEEE Trans. on Signal Processing*, vol. 89, no. 10, pp. 1959–1972, 2009.

[18] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[19] E. J. Candès, "Compressive sampling," in *Proc. of the International Congress of Mathematicians*, vol. 17, no. 4, Jan 2006, pp. 1–20.

[20] J. A. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Tran. on Information Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

[21] ——, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[22] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," *Proc. of the IEEE*, vol. 98, no. 6, pp. 948–958, 2010.

[23] B. M. Sanandaji, T. L. Vincent, M. B. Wakin, R. Tóth, and K. Poolla, "Compressive system identification of LTI and LTV ARX models," in *Proc. of the 50th IEEE Conf. on Decision and Control*, Orlando, Florida, USA, Dec. 2011.

[24] S. S. Chen, D. L. Donoho, and S. M. A, "Atomic decomposition by basis pursuit," *SIAM Journal Scientific Computing*, vol. 20, pp. 33–61, 1998.

[25] M. R. Osborne, B. Presnell, and B. Turlach, "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, vol. 20, pp. 389–403, 2000.

[26] H. Leeb and B. Pötscher, "Sparse estimators and the oracle property, or the return of hodges' estimator," *Journal of Econometrics*, vol. 142, no. 1, pp. 201–211, 2008.

[27] C. R. Rojas and H. Hjalmarsson, "SPARSEVA: Sparse estimation based on a validation criterion," in *Proc. of the 50th IEEE Conf. on Decision and Control*, Orlando, Florida, USA, Dec. 2011.

[28] M. Gevers, A. S. Bazanella, X. Bombois, and L. Mišković, "Identification and the information matrix: how to get just sufficiently rich?" *IEEE Tran. on Automatic Control*, vol. 54, no. 12, pp. 2828–2840, 2009.

[29] L. Ljung, *System Identification Toolbox, for use with Matlab*. The Mathworks Inc., 2006.