2013 European Control Conference (ECC)
July 17-19, 2013, Zürich, Switzerland.

A Sparse Estimation Technique for General Model Structures

Cristian R. Rojas, Bo Wahlberg and Håkan Hjalmarsson

Abstract— **In this paper, a general sparse estimator is proposed, based on the maximum likelihood / prediction error method (or any $\sqrt{N}$-consistent estimator). This procedure does not rely on the convexity of the cost function of the underlying estimator (in case such estimator is an M-estimator), and it provides an automatic tuning of the (implicit) regularization parameter. The idea behind the proposed method is a three step procedure, where the first step consists in a standard $\sqrt{N}$-consistent estimation, the second step seeks for the sparsest estimate in a neighborhood of the initial estimate, and the last step is a refinement based on the sparseness pattern estimated in the second step. A rigorous statistical analysis is provided, which establishes conditions for consistency, asymptotic variable selection and the so-called Oracle property. A simulation example is given to demonstrate the performance of the method.**

I. Introduction

A long standing problem in estimation is model selection. In linear regression this amounts to selecting appropriate regressors among a large set of candidate regressors. The brute force approach of comparing all possible subsets using some cross-validation method leads to combinatorial complexity. It is also problematic to analyze the statistical power of this approach.

Many approaches have been suggested to overcome these problems, such as Forward Selection [1], Forward stepwise selection, LARS (Least Angle Regression) [2] and thresholding [3]. Another class of methods that can handle all regressors at once use regularization, i.e., a penalty on the size of the parameter vector is added to the cost function. The LASSO (least absolute shrinkage and selection operator) is one of the early contributions to this field, and has been of tremendous influence[1], [5]. The LASSO is based on ℓ_1-regularization, which has close ties with the recent area of compressive sensing [6]. Other techniques, based on Bayesian formulations using sparse prior distributions for the parameters, have also been developed [7, 8]; these techniques seem to display good empirical performance, but they currently lack theoretical support.

An essential component in many of the approaches is the use of cross-validation or some information criterion, e.g. the Akaike Information Criterion (AIC) or generalized cross-validation (GCV), to determine the value of some regularization/tuning parameters. This procedure is typically computationally expensive, as it requires applying the model selection procedure for several values of the tuning parameters and then choosing the value for which the cross-validation criterion is minimum. A recent contribution, [9], has addressed this issue by proposing a tuning parameter free variant of the LASSO; see also [10].

The LASSO and other ℓ_1 techniques for model selection rely on the convexity of the unpenalized M-estimator[2]. This restriction is quite severe, and greatly constrains the set of statistical models for which such techniques can be applied, since even in areas such as system identification, the most interesting model structures such as ARMAX, Output-Error Box-Jenkins [11] give rise to non-convex prediction error cost functions. Some recent contributions [10, 12] overcome this problem by combining ℓ_1 regularization with other techniques such as Steiglitz-McBride, [11], or by relying on a matrix version of ℓ_1 regularization known as the nuclear norm, see [13, 14].

In this paper we propose a general procedure for sparse estimation which does not rely on the convexity of the cost function of the underlying estimator (in case such estimator is an M-estimator), and it provides an automatic tuning of the (implicit) regularization parameter. The idea behind the proposed method is a three step procedure, where the first step consists in a standard $\sqrt{N}$-consistent estimation, the second step seeks for the sparsest estimate in a neighborhood of the initial estimate, and the third step refines that estimate (to achieve asymptotic efficiency). The second step consists of a "linearized" version of SPARSEVA, from [9], which is a tuning parameter free ℓ_1 sparse estimator. This three step method can be seen as a variant of Söderström, Stoica and Friedlander's Indirect PEM method [15]. Some statistical properties of the proposed method are also established, namely, consistency, sparseness and the so-called Oracle property. Finally, the performance of the method is verified by a simulation example.

It is worth noting that in [16] a method similar to the one presented here was introduced. In fact, the technique in [16] can be seen as a Lagrangian variant of (3), even though, unlike the method of this paper, it depends on a regularization parameter to be tuned by the user.

The outline of the paper is as follows. In Section II the method is introduced together with the assumptions that will be used. Section III contains the main results which cover consistency, sparseness and efficiency. The method is illustrated on a numerical example in Section IV. Conclusions are provided in Section V.

Due to reasons of space, the proofs of the results of Section III have been omitted. Even though these proofs

This work was partially supported by the European Research Council under the advanced grant LEARN, contract 267381, the European Union's Seventh Framework Programme (FP7/2007-2013) AUTOPROFIT project under grant 257059, the Swedish Research Council (VR) under contract 621-2011-5890, and the Linnaeus Center ACCESS at KTH.

The authors are with the Automatic Control Laboratory and ACCESS Linnaeus Center, Electrical Engineering, KTH – Royal Institute of Technology, S-100 44 Stockholm, Sweden. Emails: {cristian.rojas|bo.wahlberg|hakan.hjalmarsson}@ee.kth.se

[1] A procedure similar to the LASSO is the *nonnegative garrote* [4]. However, as mentioned in [5], this latter method may perform poorly in overfit or highly correlated settings, while the LASSO and its variants can overcome these issues.

[2] An M-estimator is an estimator given by the solution of an optimization procedure, e.g. least squares, maximum likelihood (ML) and prediction error methods (PEM).

are similar those in [9], the interested reader can check the technical report [17] for the full details of these proofs.

Notation

$X \odot Y$ denotes the Hadamard or element-wise multiplication between two matrices X and Y of the same dimensions. Furthermore, $\|x\|_W^2 := x^T W x$ for $W = W^T > 0$ and $\|x\|_2^2 := x^T x$. $\mathrm{Cond}(A)$ is the condition number of a matrix A in the 2-norm, i.e., $\mathrm{Cond}(A) := \|A\|\|A^{-1}\|$, where $\|A\|$ denotes the maximum singular value of A. Notice that $\mathrm{Cond}(A) = \mathrm{Cond}(A^{-1}) \geq 1$. The vector containing the signs of a vector x is denoted $\mathrm{Sgn}\,[x]$. The Moore-Penrose pseudo-inverse of a matrix X is denoted $X^\dagger$.

$X_N \xrightarrow{p} X$ denotes convergence in probability, [18]. Furthermore, $A_N = O_p(B_N)$ means that, given an $\varepsilon > 0$, there exists a constant $M(\varepsilon) > 0$ and an $N_0(\varepsilon) \in \mathbb{N}$ such that for every $N \geq N_0(\varepsilon)$, $P\{|A_N| \leq M(\varepsilon)|B_N|\} \geq 1 - \varepsilon$. Similarly, $A_N = o_p(B_N)$ means that $A_N/B_N \xrightarrow{p} 0$, and $A_N \asymp_p B_N$ means that, given an $\varepsilon > 0$, there are constants $0 < m(\varepsilon) < M(\varepsilon) < \infty$ and an $N_0(\varepsilon) \in \mathbb{N}$ such that for every $N \geq N_0(\varepsilon)$, $P\{m(\varepsilon) < |A_N/B_N| < M(\varepsilon)\} \geq 1 - \varepsilon$. Also, $F(x; n)$ denotes the cumulative chi-square distribution function with n degrees of freedom.

In general, all asymptotic statements (of the form $y_N \to y$) are with respect to the number of data samples N tending to infinity.

II. Problem Statement

Consider the scalar discrete time linear time invariant stable system

$$y_t = G_0(q)u_t + H_0(q)e_t, \tag{1}$$

where u_t is a quasi-stationary (known) input signal, e_t is Gaussian white noise of zero mean and unit variance, G_0 and H_0 are stable rational functions of the forward shift operator q (i.e., $qf_t = f_{t+1}$ for every sequence f_t), with H_0 being minimum phase. Given the data $\{u_t, y_t\}_{t=1}^N$, we would like to estimate a model of G_0 and H_0 within the model structure

$$y_t = G(q;\theta)u_t + H(q;\theta)\varepsilon_t(\theta), \tag{2}$$

where $\varepsilon_t(\theta)$ is Gaussian white noise of zero mean and unit variance, and $G(q;\theta)$, $H(q;\theta)$ are stable rational functions of q parameterized by $\theta \in \mathbb{R}^n$. Here we assume that there is no undermodelling, i.e., there is a $\theta_0 \in \mathbb{R}^n$ such that $G(q;\theta_0) = G_0(q)$ and $H(q;\theta_0) = H_0(q)$, that the input is persistently exciting, and that the model structure $[G(q;\theta), H(q,\theta)]$ is globally identifiable. Furthermore, we will assume that only a small number of components of θ are non-zero, that is, we assume that θ is *sparse*. In particular, we assume without loss of generality that $\theta_0 = [\theta_0^{1T}\ \theta_0^{2T}]^T$, where $\theta_0^i \in \mathbb{R}^{n_i}$ $(i = 1, 2)$ and $\theta_0^2 = 0$. We emphasize that this is for notational convenience only; the results below hold regardless of the distribution of zeros in θ_0. Our goal then involves determining the location of the non-zero entries of θ, which is essentially a *model structure selection problem.*

A standard method for estimating θ (without imposing sparseness) is the Prediction Error Method [11] (PEM), given by

$$\hat{\theta}_N^{\text{PEM}} := \arg\min_{\theta \in \mathbb{R}^n} V_N(\theta);$$

$$V_N(\theta) := \frac{1}{N}\sum_{t=1}^{N} \varepsilon_t^2(\theta), \quad \varepsilon_t(\theta) := \frac{1}{H(q;\theta)}[y_t - G(q;\theta)u_t].$$

Unfortunately, for many model structures, such as ARMAX, Output-Error or Box-Jenkins [11], the cost function V_N minimized by PEM is non-convex, and it may have several local minima, which implies that the numerical computation of $\hat{\theta}_N^{\text{PEM}}$ can be a challenging problem. On the other hand, we can content ourselves with an asymptotically efficient estimator $\hat{\theta}_N$, that is, one whose covariance matrix $\mathrm{Cov}(\hat{\theta}_N)$ satisfies $\lim_{N\to\infty} N\mathrm{Cov}(\hat{\theta}_N) = I_F^{-1}(\theta_0) =: P$, where $I_F(\theta)$ is the normalized (by N) Fisher information matrix [19] for the model structure (2). In fact, it can be shown, under mild conditions, that given an initial $\sqrt{N}$-consistent estimator $\tilde{\theta}_N$ (based, e.g., on instrumental variables or subspace identification methods), an application of only one step of the Gauss-Newton method on the cost function of PEM (using $\tilde{\theta}_N$ as the initial condition) gives an asymptotically efficient estimator [20, Section 5.5.2]. For this reason, we will assume in the sequel that we have an asymptotically efficient estimator of θ, say, $\hat{\theta}_N^{\text{init}}$, and furthermore, a consistent estimator, $\hat{P}_N$, of its normalized (asymptotic) covariance, P.

In order to impose the sparseness constraint on the estimate of θ, we propose the following algorithm, based on Adaptive SPARSEVA, a technique developed in [9]:

1) Obtain an asymptotically efficient estimator of θ, $\hat{\theta}_N^{\text{init}}$, and a consistent estimator $\hat{P}_N$.
2) Solve the optimization program:

$$\begin{array}{ll} \min\limits_{\theta \in \mathbb{R}^n} & \|w_N \odot \theta\|_1 \\ \text{s.t.} & (\theta - \hat{\theta}_N^{\text{init}})^T \hat{P}_N^{-1} (\theta - \hat{\theta}_N^{\text{init}}) \leq \delta_N, \end{array} \tag{3}$$

   where $w_N := [1/|(\hat{\theta}_N^{\text{init}})_1| \ \cdots \ 1/|(\hat{\theta}_N^{\text{init}})_n|]^T$ and $\delta_N > 0$. Notice that, under mild assumptions on the estimator $\hat{\theta}_N^{\text{init}}$ (for example, if it is a maximum likelihood estimator, if the distribution of the data is absolutely continuous and it depends smoothly on θ), the weights in w_N are finite with probability 1. The choice of δ_N will be discussed later. Let us denote the solution of (3) as $\hat{\theta}_N$.
3) Finally, re-estimate the non-zero elements of $\hat{\theta}_N$. More precisely, apply one step of the Gauss-Newton method to $V_N(\theta)$ starting from $\hat{\theta}_N^{\text{init}}$ and forcing the components of θ corresponding to the zero entries of $\hat{\theta}_N$ to remain equal to zero. This means solving

$$\begin{array}{ll} \min\limits_{\theta \in \mathbb{R}^n} & (\theta - \hat{\theta}_N^{\text{init}})^T \hat{P}_N^{-1} (\theta - \hat{\theta}_N^{\text{init}}) \\ \text{s.t.} & \theta_i = 0, \quad \text{for all } i = 1, \ldots, N \text{ s.t. } (\hat{\theta}_N)_i = 0. \end{array}$$

   Let us denote this refined estimator as $\hat{\theta}_N^{RE}$.

The intuition behind this procedure is as follows: The estimate $\hat{\theta}_N^{\text{init}}$ is asymptotically efficient, but not necessarily sparse. Hence, if θ_0, the true value of θ, is sparse, then it should lie in a neighborhood of $\hat{\theta}_N^{\text{init}}$. One way to find it is to set up a confidence ellipsoid around $\hat{\theta}_N^{\text{init}}$, and to determine the "sparsest" θ in such region; this is essentially

done by solving (3), which relies on the ℓ_1 norm to enforce sparseness. Now, even though the solution of (3) may have the right amount of sparseness (in other words, it may have correctly determined which components of θ_0 are zero), it does not necessarily provide a good estimate of the non-zero entries of θ_0. This last problem can be remedied by using $\hat{\theta}_N^{\text{init}}$ as an initial condition for a Gauss-Newton iteration, and using the sparseness structure of $\hat{\theta}_N$ as a constraint, which would then yield an asymptotically efficient estimate of θ_0 (conditioned on the event that the sparseness of θ_0 has been correctly determined).

In order to choose δ_N, it is first important to notice that $N(\theta - \hat{\theta}_N^{\text{init}})^T \hat{P}_N^{-1}(\theta - \hat{\theta}_N^{\text{init}}) \in \mathbf{AsN}(0, I)$. This means that as $N \to \infty$, the constraint set in (3) can be interpreted as a *confidence region* of level $F(N\delta_N; n)$. In close connection to [21], where several model selection criteria are related to the F-test with a suitable choice of the significance level, it can be seen that in order to correctly detect the true location of the zeros of θ_0, the confidence region should include θ_0 (or at least a θ with the same sparseness pattern as θ_0). This condition holds if and only if $N\delta_N \to \infty$. These issues are studied in more rigor in Section III. Two typical choices for δ_N are:

- $\delta_N = 2n/N$.
- $\delta_N = n\ln(N)/N$.

The first choice is motivated by the AIC criterion, while the second one is related to the BIC/MDL criterion [11].

Remark 2.1: In case it is desired to impose sparseness only on a subset of the parameters in θ, (3) can be modified so that only that sparse component of θ is considered in the cost function and the constraint. In addition, $\hat{P}_N^{-1}$ has to be replaced by the inverse of the covariance matrix of the sparse subset of θ (which is obtained simply by removing the rows and columns of $\hat{P}_N$ corresponding to the non-sparse part of θ, and taking the inverse of such matrix).

Remark 2.2: Notice that the method proposed above is applicable under much more general conditions than those stated (i.e., linear time invariant systems). In fact, it can in principle be used for general statistical models, under mild conditions guaranteeing consistency and asymptotic normality of maximum likelihood estimators. However, for simplicity of presentation, we have kept a more restricted and familiar setting for (linear) system identification.

III. Theoretical Analysis

In this section we present the main technical results of the proposed method. Notice that for the asymptotic results it is assumed that the model structure is kept fixed as $N \to \infty$, i.e., n is fixed. The proofs of these results are similar to those of A-SPARSEVA in [9], but for completeness they are presented in the Appendices.

A. Consistency

A first concern regarding the proposed algorithm is whether (and under which conditions) $\hat{\theta}_N$ converges in probability to θ_0. This is addressed in the following theorem.

Theorem 3.1 (Consistency): Under the stated assumptions, and[3] $\theta_0 \neq 0$, we have that $\hat{\theta}_N$ is consistent in probability (i.e. $\hat{\theta}_N \xrightarrow{p} \theta_0$ and $\hat{\theta}_N^{RE} \xrightarrow{p} \theta_0$) if and only if $\delta_N \to 0$. In particular, $\|\hat{\theta}_N - \theta_0\|_2 = O_p(N^{-1/2} + \sqrt{\delta_N})$.

The properties of the refined estimator $\hat{\theta}_N^{RE}$ will be discussed later (in relation to the sparseness and Oracle properties).

Corollary 3.1 (Exact order of consistency): Subject to the assumptions of Theorem 3.1, if $\delta_N \to 0$ but $N\delta_N \to \infty$, then $\|\hat{\theta}_N - \theta_0\|_2 \asymp_p \sqrt{\delta_N}$.

B. Sparseness

We now establish the exact conditions on δ_N for $\hat{\theta}_N$ to generate sparse estimates.

Theorem 3.2 (Sparseness): Consider the stated assumptions, and in addition $\delta_N \to 0$ and $\theta_0 \neq 0$. Then, $\hat{\theta}_N$ satisfies the *sparseness property* (i.e., $\hat{\theta}_N = [(\hat{\theta}_N^1)^T \ (\hat{\theta}_N^2)^T]^T$, with $\hat{\theta}_N^i \in \mathbb{R}^{n_i}$ $(i = 1, 2)$, where $P\{\hat{\theta}_N^2 = 0\} \to 1$) if $N\delta_N \to \infty$. If $N\delta_N \to \infty$ does not hold, $\hat{\theta}_N$ does not have the sparseness property.

Notice that, according to Theorem 3.2, the BIC choice $\delta_N = n\ln(N)/N$ gives sparseness, while the AIC choice $\delta_N = 2n/N$ does not.

C. Oracle property

From the preceding results, $\hat{\theta}_N$ has the sparseness property if and only if δ_N is chosen such that $\delta_N \to 0$ and $N\delta_N \to \infty$. On the other hand, by Corollary 3.1, such a choice of δ_N gives rise to a non efficient estimator (since the order of convergence of $\hat{\theta}_N$ to θ_0 would be $\sqrt{\delta_N}$, according to Corollary 3.1, strictly larger than $N^{-1/2}$). The third step of the proposed procedure can overcome this efficiency-sparseness tradeoff by refining the estimator. The next result shows that the estimator obtained from the third step of our method is asymptotically normal and efficient.

Theorem 3.3 (The Oracle property): Consider the assumptions in Theorem 3.2 and that $N\delta_N \to \infty$. Then

$$\sqrt{N}(\hat{\theta}_N^{RE} - \theta_0) \in \mathbf{AsN}(0, M^\dagger)$$

where M is the information matrix when it is known which elements of θ_0 are zero.

IV. Numerical Example

In this section we give a simple simulation example, based on the problem of acoustic echo estimation [22] with colored (autorregressive) measurement noise, to illustrate the use of the proposed method.

A. Preliminaries

Firstly, we present the model for the example and then derive an initial asymptotically efficient estimator. Consider the model

$$y_t = \sum_{k=1}^{K} b_k u_{t-k} + v_t, \tag{4}$$

$$v_t = \frac{1}{D(q)} e_t, \quad D(q) = \sum_{m=0}^{d} d_m q^{-m}, \quad d_0 := 1,$$

where $\{b_k\}_{k=1}^K$ and $\{d_m\}_{m=1}^d$ are real unknown parameters, e_t is Gaussian white noise of variance σ^2, and u_t is a known input signal. It is known *a priori* that all of the

[3] Notice that the assumption $\theta_0 \neq 0$ is considered merely to avoid a special case in the proof. In fact, all theorems in Section III hold if $\theta_0 = 0$.

parameters b_k are zero except for a small number. Notice that (4) does not fit into a standard linear regression framework, hence Lasso and other ℓ_1-penalized convex estimators cannot be directly applied without losing statistical efficiency. In order to estimate these parameters using the method of Section II, we need an initial asymptotically efficient non-sparse estimator. To this end, notice that from (4)

$$\begin{aligned} D(q)y_t &= \sum_{m=0}^{d}\sum_{k=1}^{K} d_m b_k u_{t-k-m} + e_t \\ &= \sum_{T=1}^{d+K} \underbrace{\left(\sum_{m=\max\{0,T-K\}}^{T-\max\{1,T-d\}} d_m b_{T-m} \right)}_{=:\tilde{b}_T} u_{t-T} + e_t \\ &= \sum_{T=1}^{d+K} \tilde{b}_T u_{t-T} + e_t = \phi_t^T \tilde{B} + e_t, \end{aligned} \tag{5}$$

where $\phi_t := [u_{t-1} \ \cdots \ u_{t-d-K}]^T \in \mathbb{R}^{(d+K)\times 1}$ and $\tilde{B} := [\tilde{b}_1 \ \cdots \ \tilde{b}_{d+K}]^T \in \mathbb{R}^{(d+K)\times 1}$. We can now collect the N data samples in vectors[4] $Y_{-i} := [y_{1-i} \ \cdots \ y_{N-i}]^T \in \mathbb{R}^{N\times 1}$ and $E := [e_1 \ \cdots \ e_N]^T \in \mathbb{R}^{N\times 1}$, and define $\Phi := [\phi_1 \ \cdots \ \phi_N]^T \in \mathbb{R}^{N\times(d+K)}$, so that (5) can be written as

$$Y_0 = [Y_{-1} \ \cdots \ Y_{-d} \ \Phi][d_1 \ \cdots \ d_d \ \mathrm{vec}(\tilde{B})^T]^T + E. \tag{6}$$

By applying least squares to (6), we can obtain initial estimates of $d_1, \ldots, d_d, \tilde{b}_1, \ldots, \tilde{b}_{d+K}$. Let us denote by $\hat{D}$ and $\hat{\tilde{B}}$ the vectors of estimates of these quantities. Notice that the parameters $\tilde{b}_1, \ldots, \tilde{b}_{d+K}$ are related to $b_1, \ldots, b_K$ through the mapping $\tilde{B} = T_D B$, where $B := [b_1 \ \cdots \ b_K]^T \in \mathbb{R}^{K\times 1}$ and

$$T_D := \begin{bmatrix} d_0 & \cdots & d_d & & & & \\ & d_0 & & d_d & & 0 & \\ & & \ddots & & \ddots & & \\ & 0 & & d_0 & & d_d & \\ & & & & d_0 & \cdots & d_d \end{bmatrix}^T \in \mathbb{R}^{(d+K)\times K}.$$

Based on these equations, an initial $\sqrt{N}$-consistent estimate of $b_1, \ldots, b_K$ can be obtained e.g. from[5]

$$\tilde{B} = \arg\min_U \|\hat{\tilde{B}} - T_{\hat{D}} B\|_2^2 = (T_{\hat{D}}^T T_{\hat{D}})^{-1} T_{\hat{D}}^T \hat{\tilde{B}}.$$

In order to obtain initial asymptotically efficient estimates of $d_1, \ldots, d_d, b_1, \ldots, b_{d+K}$, we can use the indirect PEM approach [15], which consists in applying one Gauss-Newton iteration to $\hat{D}$ and $\hat{B}$:

$$\begin{bmatrix} \hat{D}^{\text{init}} \\ \hat{B}^{\text{init}} \end{bmatrix} := \begin{bmatrix} \hat{D} \\ \hat{B} \end{bmatrix} + (G\tilde{P}^{-1}G^T)^{-1} G\tilde{P}^{-1} \begin{bmatrix} 0 \\ \hat{\tilde{B}} - T_{\hat{D}}\hat{B} \end{bmatrix},$$

where $\hat{D}^{\text{init}}$ and $\hat{B}^{\text{init}}$ denote asymptotically efficient estimators of D and B, and

$$G := \left.\frac{\partial [D^T \ \tilde{B}^T]}{\partial \begin{bmatrix} D \\ B \end{bmatrix}}\right|_{D=\hat{D},\, B=\hat{B}} = \begin{bmatrix} I_{d\times d} & T_{\hat{B}}^T \\ 0_{K\times d} & T_{\hat{D}}^T \end{bmatrix},$$

$$T_B := \begin{bmatrix} 0 & b_1 & \cdots & b_K & & & & \\ & 0 & b_1 & & b_K & & 0 & \\ & & \ddots & \ddots & & \ddots & & \\ & 0 & & 0 & b_1 & & b_K & \\ & & & & 0 & b_1 & \cdots & b_K \end{bmatrix}^T$$

$$\tilde{P}^{-1} := N^{-1}[Y_{-1} \ \cdots \ Y_{-d} \ \Phi]^T [Y_{-1} \ \cdots \ Y_{-m} \ \Phi],$$

where $T_B \in \mathbb{R}^{(d+K)\times d}$. Furthermore, the asymptotic covariance matrix of $[\hat{D}^{\text{init}\,T} \ \hat{B}^{\text{init}\,T}]^T$ is $\hat{P}^{\text{init}} = (G\tilde{P}^{-1}G^T)^{-1}$.

B. Simulations

Consider the system (4) where $B_0 = [3 \ 1.5 \ 0 \ 0 \ 2 \ 0 \ 1 \ 0 \ 0 \ 0.4]^T$, $D_0(q) = 1 - 1.7q^{-1} + 0.72q^{-2}$ and $\sigma = 0.3$. This means that the system is of the form (1), with

$$\begin{aligned} G_0(q) &= 3q^{-1} + 1.5q^{-2} + 2q^{-5} + q^{-7} + 0.4q^{-10} \\ H_0(q) &= \frac{0.3}{1 - 1.7q^{-1} + 0.72q^{-2}}. \end{aligned}$$

The input signal u_t is generated as Gaussian white noise of zero mean and unit variance. Since the impulse response coefficients of G are sparse, our goal is to use the method proposed in Section II to estimate such coefficients. To this end, we will choose $\delta_N = K\ln(N)/N$, resembling the BIC criterion, for which the sparseness pattern of the b_k's can be recovered (asymptotically in N), according to Theorem 3.2.

Figures 1 and 2 present the results of 100 Monte Carlo simulations for each value of N. Here, "normalized MSE" refers to the (normalized by N) total mean square error (MSE) of the parameter estimators, i.e., $N\|\hat{B} - B_0\|_2^2$, where $\hat{\theta}$ corresponds to the respective estimator. From Figure 1, we can notice that while the asymptotically efficient estimator $\hat{B}^{\text{init}}$ represents a marginal improvement over the initial estimator $\tilde{B}$, the "refined sparse PEM" estimator $\hat{B}_N^{RE}$ has a much smaller MSE, due to its successful detection of the sparseness pattern of B_0 (as shown in Figure 2). The *unrefined* "sparse PEM" estimator $\hat{B}_N$ has a much higher normalized MSE than the other estimators, which even increases (logarithmically) with N. This is because such estimator lies by construction on the boundary of the confidence ellipsoid used in (3), whose linear size is proportional to $N\delta_N = K\ln(N)$.

V. Conclusions

In this paper we have proposed a new method for sparse estimation, which can handle general model structures (i.e., not restricted to linear parameterizations). This estimator consists of three steps: the first step is a standard asymptotically efficient estimator (for which several techniques are available), the second step imposes sparseness via ℓ_1 minimization over a confidence ellipsoid, and the final step refines such estimate. The supporting theory and a simple

[4]For simplicity of presentation, we have assumed $y_t = 0$ for $t < 0$. In practice, we can also eliminate the first d components of Y_{-i}. The difference between these two approaches becomes negligible as $N \to \infty$.

[5]In this section, we take as notation that expressions such as $T_{\hat{D}}$ correspond to T_D where all entries related to D are replaced by their corresponding estimates from $\hat{D}$.

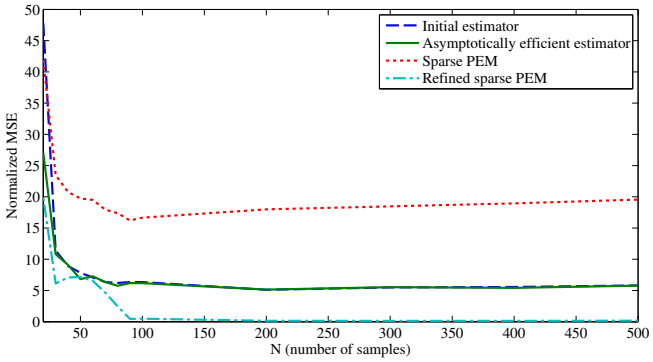

Fig. 1. Normalized MSE of the proposed estimator of B ("Refined sparse PEM", $\hat{B}_N^{RE}$), and the intermediate estimators used to construct it. Here "Initial estimator", "Asymptotically efficient estimator" and "Sparse PEM" correspond to $\tilde{B}$, $\hat{B}^{\text{init}}$ and $\hat{B}_N$, respectively, following the notation of Sections IV-A and II.

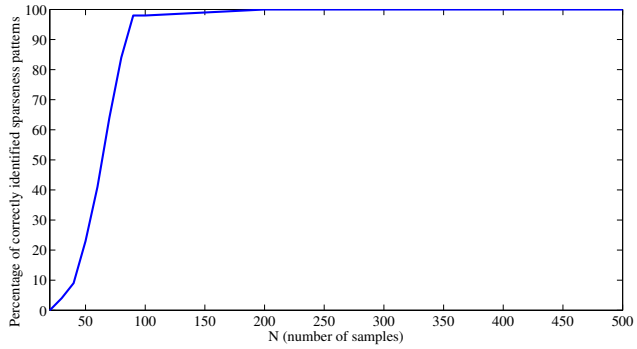

Fig. 2. Percentage of the 100 Monte Carlo simulations for which the proposed method (with $\delta_N = K\ln(N)/N$) has determined correctly which components of B are zero and which are not.

example presented here exhibit the potential and simplicity of the method for sparse estimation for almost arbitrary model structures.

REFERENCES

[1] S. Weisberg, *Applied Linear Regression*. New York: Wiley, 1980.

[2] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–451, 2004.

[3] D. Donoho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[4] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37(4), pp. 373–384, 1995.

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

[6] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[7] D. P. Wipf and S. Nagarajan, "A new view of automatic relevance determination," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. MIT Press, 2008.

[8] A. Chiuso and G. Pillonetto, "A Bayesian approach to sparse dynamic network identification," *Automatica*, vol. 48(8), pp. 1553–1565, 2012.

[9] C. R. Rojas and H. Hjalmarsson, "Sparse estimation based on a validation criterion," in *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC11)*, Orlando, USA, 2011.

[10] R. Tóth, H. Hjalmarsson, and C. R. Rojas, "Sparse estimation of rational dynamic models," in *Proceedings of the 16th IFAC Symposium on System Identification (SYSID 2012) (accepted for publication)*, 2012.

[11] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[12] H. Hjalmarsson, J. S. Welsh, and C. R. Rojas, "Identification of Box-Jenkins models using structured ARX models and nuclear norm relaxation," in *Proceedings of the 16th IFAC Symposium on System Identification (SYSID 2012) (accepted for publication)*, 2012.

[13] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the American Control Conference (ACC'01)*, vol. 6, 2001, pp. 4734–4739.

[14] Z. Liu and L. Vandenberghe, "Interior-pont method for nuclear norm approximation with application to system identification," *SIAM Journal of Matrix Analysis and Applications*, vol. 31(3), pp. 1235–1256, 2009.

[15] T. Söderström, P. Stoica, and B. Friedlander, "An indirect prediction error method for system identification," *Automatica*, vol. 27(1), pp. 183–188, 1991.

[16] H. Wang and C. Leng, "Unified LASSO estimation by least squares approximation," *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 1039–1048, 2007.

[17] C. R. Rojas, B. Wahlberg, and H. Hjalmarsson, "A sparse estimation technique for general model structures," *http://www.ee.kth.se/˜crro/sparsePEM.pdf*, 2013, technical Report.

[18] E. L. Lehmann, *Elements of Large-Sample Theory*. Springer, 1999.

[19] G. C. Goodwin and R. L. Payne, *Dynamic System Identification: Experiment Design and Data Analysis*. New York: Academic Press, 1977.

[20] W. A. Fuller, *Introduction to Statistical Time Series, 2nd Edition*. John Wiley & Sons, 1996.

[21] T. Söderström, "On model structure testing in system identification," *International Journal of Control*, vol. 26(1), pp. 1–18, 1977.

[22] S. F. Cotter and B. D. Rao, "Sparse channel estimation via matching pursuit with application to equalization," *IEEE Transactions on Communications*, vol. 50(3), pp. 374–377, 2002.