Digital Signal Processing 54 (2016) 27–34

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

On a class of parameters estimators in linear models dominating the least squares one

Piero Barone [a,*], Isabella Lari [b]

[a] Istituto per le Applicazioni del Calcolo "M. Picone", C.N.R., Via dei Taurini 19, 00185 Rome, Italy
[b] Dipartimento di Scienze Statistiche, Sapienza Universita' di Roma, Piazzale Aldo Moro 5, 00185 Rome, Italy

ARTICLE INFO

Article history:
Available online 8 April 2016

Keywords:
Linear model
Mean square error
Biased estimates
Noise model
l_1 norm minimization
Ill-posed inverse problems

ABSTRACT

The estimation of parameters in a linear model is considered under the hypothesis that the noise, with finite second order statistics, can be represented by random coefficients in a given deterministic basis. An extended underdetermined design matrix is then formed, and the estimator of the extended parameters with minimum l_1 norm is computed. It is proved that, if the noise variance is larger than a threshold, which depends on the unknown parameters and on the extended design matrix, then the proposed estimator of the original parameters dominates the least-squares estimator, in the sense of the mean square error. A small simulation illustrates its behavior. Moreover it is shown experimentally that it can be convenient, even if the design matrix is not known but only an estimate can be used. Furthermore the noise basis can eventually be used to introduce some prior information in the estimation process. These points are illustrated in a simulation by using the proposed estimator for solving a difficult inverse ill-posed problem, related to the complex moments of an atomic complex measure.

Introduction

Linear models are ubiquitous in applied sciences. Parameters estimation methods have been developed since a long time ago. In order to motivate the approach that we are proposing, we make some considerations on parameters estimation in linear models related to our purpose. Denoting random quantities by bold characters, let us consider the model

$$\underline{\mathbf{d}} = V\underline{\xi} + \underline{\boldsymbol{\epsilon}}, \;\; V \in \mathbb{C}^{n\times p}, \; n \geq p, \; \text{rank}(V) = p, \; \underline{\xi} \in \mathbb{C}^p,$$

where $\underline{\mathbf{d}}$ is an n-variate complex random vector representing the measured data and $\underline{\boldsymbol{\epsilon}}$ is a n-variate zero mean complex random vector with variance σ^2 representing the measuring error. The design matrix V is assumed to be ill-conditioned w.r.t. the inversion, i.e. the ratio of its largest to the smallest singular value is large.

The classical least squares (LS) estimator is

$$\underline{\boldsymbol{\xi}}_{LS} = V^+\underline{\mathbf{d}} = (V^TV)^{-1}V^T\underline{\mathbf{d}},$$

where T denotes transposition plus conjugation, and $+$ denotes generalized inversion. The first and second order statistics of $\underline{\boldsymbol{\xi}}_{LS}$ are

$$E[\underline{\boldsymbol{\xi}}_{LS}] = \underline{\xi}, \;\; \text{cov}[\underline{\boldsymbol{\xi}}_{LS}] = V^+\text{cov}[\underline{\boldsymbol{\epsilon}}]V^{+^T}.$$

Therefore the LS estimator is not distorted and, when the error is independent and identically distributed with variance σ^2, its covariance and mean square error (MSE) reduces to

$$MSE_{LS} = \text{tr}(\text{cov}[\underline{\boldsymbol{\xi}}_{LS}]) = \sigma^2\text{tr}((V^TV)^{-1}).$$

In the following this setup will be assumed. When the design matrix is ill-conditioned, MSE_{LS} can be quite large. In many instances this can be a serious problem because of the consequent instability of the estimates. It is therefore reasonable to allow some bias in the estimators in order to reduce their variability measured by the MSE. Several methods are reported in the literature which, from a mathematical point of view, reduce either to solve a min-max problem (e.g., [13,5]) or to solve a l_1 minimization with quadratic constraints problem or, equivalently, a quadratic minimization with l_1 constraints (e.g., [14,11,8]). Estimators in the two classes are quite distinct and it can be proved that methods in one class do not dominate uniformly methods in the other one [12].

In many applications, the parameter vector $\underline{\xi}$ to be estimated represents some well defined object about which much a priori information is available. This motivated the introduction of regularization methods which enforces a parameter estimate with expected properties by solving a modified problem, e.g., of the form

* Corresponding author.
E-mail addresses: p.barone@iac.cnr.it, piero.barone@gmail.com (P. Barone), isabella.lari@uniroma1.it (I. Lari).

http://dx.doi.org/10.1016/j.dsp.2016.04.001
1051-2004/

$$\hat{\underline{\xi}} = \text{argmin}_{\underline{x}} \left(\|\underline{d} - V\underline{x}\|^2 + \lambda f(\underline{x}) \right),$$

where $f(\underline{x}) \geq 0$ is a regularization function which represents the prior information and $\lambda > 0$ is an hyperparameter balancing the fit to the data and the prior information. In a stochastic environment, the Bayes paradigm implements the same idea in a more general form. Given a prior distribution of the parameters and a likelihood, a function proportional to the posterior is used to get estimators, either by solving an optimization problem, or simply by sampling from the posterior. This last approach is able to cope with problems of huge dimension (see, e.g., [10] where the Markov Chain Monte Carlo approach is described).

In this work, a different approach is pursued with the same aim. The basic observation is that sometimes we are not able to characterize the parameters but we are able to characterize the noise quite well. As an example of this situation, we quote the complex exponentials approximation problem [2,4], where it is well known that, under a suitable coordinate transformation, the noise clusters around the unit circle in the complex plane, but for some gaps, accordingly to an equilibrium measure induced by a logarithmic potential [1]. Moreover, this behavior is quite general (universal), i.e. it does not depend on the specific distribution of the noise [3,6]. The idea is then to consider a model for the noise

$$\underline{\boldsymbol{\epsilon}} = V_e \underline{\boldsymbol{\eta}}, \quad V_e \in \mathbb{C}^{n \times (m-p)}, \; m > p, \tag{1}$$

where the matrix V_e is assigned on the basis of the assumed information about the noise and $\underline{\boldsymbol{\eta}}$ is a random complex vector of noise-related parameters to be estimated. We can then consider an extended model

$$\underline{\mathbf{d}} = V\underline{\xi} + V_e\underline{\boldsymbol{\eta}} = [V|V_e][\underline{\xi}^T|\underline{\boldsymbol{\eta}}^T]^T = \mathcal{M}\underline{\mathbf{x}}, \tag{2}$$

where

$$\mathcal{M} \in \mathbb{C}^{n \times m}, \; n \leq m, \; \text{rank}(\mathcal{M}) \geq p$$

and $\mathcal{M}$ is underdetermined if $m > n$. We have now a problem similar to a compressed sensing problem [11,8] with the important simplification that we know which are the noise related components in the extended design matrix $\mathcal{M}$. In order to exploit this similarity, we make use of the real isomorph transformation to reformulate the problem in real variables

$$\underline{x} \in \mathbb{C}^r \to \begin{bmatrix} \Re\underline{x} \\ \Im\underline{x} \end{bmatrix} \in \mathbb{R}^{2r},$$

$$X \in \mathbb{C}^{r \times s} \to \begin{bmatrix} \Re X & -\Im X \\ \Im X & \Re X \end{bmatrix} \in \mathbb{R}^{2r \times 2s}.$$

Hence, in the following only the real case will be discussed, but in the last section where we implicitly assume that the real isomorph transformation is used. We can then consider the estimator given by

$$\begin{cases} \hat{\underline{\mathbf{x}}} = \text{argmin}_{\underline{\mathbf{x}}} \|\underline{\mathbf{x}}\|_1 \\ \underline{\mathbf{d}} = \mathcal{M}\underline{\mathbf{x}} \end{cases}, \quad \text{where } \hat{\underline{\mathbf{x}}} = \begin{bmatrix} \underline{\boldsymbol{\xi}}_D \\ \hat{\underline{\boldsymbol{\eta}}} \end{bmatrix} \tag{3}$$

and find conditions on $\underline{\xi}$, V_e and σ^2 such that

$$MSE_D = E[\|\underline{\xi} - \underline{\boldsymbol{\xi}}_D\|_2^2] < MSE_{LS}.$$

We notice that, by introducing Lagrange multipliers, the problem above is equivalent to a regularization problem with a special regularization function given by the l_1 norm of the extended unknown vector.

In the following, an explicit expression of the solution of problem (3) is derived which allows to study conditions under which the new estimator dominates the LS one. These conditions depend on unknown quantities and therefore are of little practical relevance but in specific cases. However, we can prove existence theorems and get hints on the choice of matrix V_e, if not "a priori" known. Moreover, it turns out that the proposed method is convenient if the noise variance is large enough and this is confirmed by the reported numerical experiments. The paper is organized as follows. In section 1, an explicit form of the estimator is provided. In section 2, conditions on $\underline{\xi}$, V, V_e and σ^2 are derived in order to prove that there exist sets of parameters such that the proposed estimator dominates the LS one, and an easy-to-verify condition in a specific case is discussed. In section 3, a small simulation related to the difficult problem of complex exponential approximation is performed to illustrate the advantages of the considered estimator; moreover, a comparison with a blind minimax method described in [5] is provided.

1. Explicit form of the estimator

In order to get an explicit form of the estimator $\underline{\boldsymbol{\xi}}_D$, let us consider the case $m = n + p$. We first state the following result.

Lemma 1. *If $B \in \mathbb{R}^{n \times p}$ has rank p, the problem*

$$z^* = \min_{\underline{x}} \sum_{i=1}^{n} |(\underline{b} - B\underline{x})_i| \tag{4}$$

has at least one solution of the form

$$\hat{\underline{x}} = B_p^{-1}\underline{b}_p,$$

where B_p is a non-singular submatrix of B of order p and $\underline{b}_p$ is the corresponding subvector of $\underline{b}$.

Proof. Problem (4) admits a finite optimal solution, since it is a minimization problem and $\sum_{i=1}^{n} |(\underline{b} - B\underline{x})_i| \geq 0$ for all $\underline{x} \in \mathbb{R}^p$. Let $\underline{x}^*$ be such an optimal solution and let M^+, M^- be the partition of $\{1, \ldots, n\}$ such that:

$$\begin{aligned} &(\underline{b} - B\underline{x}^*)_i \geq 0 \quad && i \in M^+ \\ &(-\underline{b} + B\underline{x}^*)_i > 0 \quad && i \in M^-. \end{aligned}$$

Consider the following Linear Program based on the partition M^+, M^-:

$$\begin{aligned} w^* \;=\; & \min_{\underline{x}} \textstyle\sum_{i \in M^+} (\underline{b} - B\underline{x})_i + \sum_{i \in M^-} (-\underline{b} + B\underline{x})_i \\ & (\underline{b} - B\underline{x})_i \geq 0, \quad i \in M^+ \\ & (-\underline{b} + B\underline{x})_i \geq 0, \quad i \in M^-. \end{aligned} \tag{5}$$

Problem (4) is a relaxation of problem (5) and, in particular, for all feasible solutions of (5), the objective function of (4) is equal to the objective function of (5); furthermore, the optimal solution $\underline{x}^*$ to (4) is feasible for (5). It follows that $\underline{x}^*$ is optimal also for (5) and $z^* = w^*$. Hence, any optimal solution to (5) is optimal also for (4).

Since problem (5) has a finite optimal solution and B has rank p, for the fundamental theorem of Linear Programming (see, e.g., [9]), there exists at least an optimal basic feasible solution to (5), i.e. an optimal solution of the form

$$\hat{\underline{x}} = B_p^{-1}\underline{b}_p,$$

where B_p is a non-singular submatrix of B of order p and $\underline{b}_p$ is the corresponding subvector of $\underline{b}$. Hence, by the above results, also problem (4) has at least an optimal solution having this form. □

The following proposition holds.