WEIGHTED AND STRUCTURED SPARSE TOTAL LEAST-SQUARES FOR PERTURBED COMPRESSIVE SAMPLING

Hao Zhu[†]*, Georgios B. Giannakis*[†]*, and Geert Leus*[*]

[†]Dept. of ECE, University of Minnesota, 200 Union St. SE, Minneapolis, MN 55455, USA
[*]Delft University of Technology, Fac. EEMCS, Mekelweg 4, 2628CD, Delft, Netherlands

ABSTRACT

Solving linear regression problems based on the total least-squares (TLS) criterion has well-documented merits in various applications, where perturbations appear both in the data vector as well as in the regression matrix. Weighted and structured generalizations of the TLS approach are further motivated in several signal processing and system identification related problems. On the other hand, modern compressive sampling and variable selection algorithms account for perturbations of the data vector, but not those affecting the regression matrix. The present paper addresses also the latter by introducing a weighted and structured sparse (S-) TLS formulation to exploit a priori knowledge on both types of perturbations, and on the sparsity of the unknown vector. The resultant novel approach is further able to cope with sparse, under-determined errors-in-variables models with structured and correlated perturbations, while allowing for efficient sub-optimum solvers. Simulated tests demonstrate the approach, and especially its ability to reliably recover the support of unknown sparse vectors.

Index Terms— Total least-squares, errors-in-variables models, sparsity, coordinate descent.

1. INTRODUCTION

Sparsity is a property possessed by many signal vectors either naturally, or, after projecting them over appropriate bases. It has been exploited for a while in numerical linear algebra, statistics, and signal processing, but renewed interest emerged recently because sparsity plays a key role in modern compressive sampling (CS) theory and applications; see e.g., [1]. Using the basis pursuit (BP) approach [2], CS can cope with noisy data when fitting parsimonious signal representations – a task of major importance for signal compression and feature extraction. The Lagrangian form of BP is also popular in statistics for fitting sparse linear regression models, using the so-termed least-absolute shrinkage and selection operator (Lasso); see e.g., [7], [4], and references thereof. However, existing CS, BP, and Lasso-based approaches do not account for perturbations present in the matrix of equations, which in the BP (respectively Lasso) circles is known as the representation basis or dictionary (correspondingly regression) matrix.

Such perturbations appear when there is a *mismatch* between the adopted basis matrix and the actual but *unknown* one – a performance-critical issue in e.g., sparsity-exploiting approaches to localization, time delay, and Doppler estimation in communications, radar, and sonar applications; see e.g., [9]. Performance analysis of CS and BP approaches for the partially-perturbed linear model with perturbations only in the basis matrix, as well as for the fully-perturbed one with perturbations present also in the measurements, was pursued recently in [5] and [3]. But devising a systematic approach to reconstructing sparse vectors under either type of perturbed models was left open.

Interestingly, for *non-sparse* over-determined linear systems, such an approach is available within the framework of total least-squares (TLS), the basic generalization of LS tailored for fitting fully-perturbed linear models [6]. For fully-perturbed, under-determined systems with sparse unknown vectors, a universal sparse (S-)TLS approach was reported recently, but without accounting for possibly available a priori information on the underlying perturbations [9]. Since structural or statistical information on the perturbations is often available, the goal of this paper is to account for it, and thereby generalize the S-TLS framework. Specifically, algorithms will be developed to solve the weighted and structured S-TLS problem, which is non-convex and thus challenging. Iterative efficient solvers will be asserted convergent to a stationary point. Maximum a posteriori optimality will be established for the well-known errors-in-variables (EIV) model. And pertinent analytical claims will be corroborated via simulations.

Notation: Upper (lower) bold face letters are used throughout to denote matrices (column vectors); $(\cdot)^T$ denotes transposition; $(\cdot)^\dagger$ the matrix pseudo-inverse; $\mathrm{vec}(\cdot)$ the column-wise matrix vectorization; $\mathrm{bdiag}(\cdot)$ the matrix block diagonalization; $\otimes$ the Kronecker product; $\mathbf{0}_m$ the $m \times 1$ vector of all zeros; $\mathbf{I}$ the identity matrix of appropriate dimensions; $\|\cdot\|_F$ the Frobenius norm; $\|\cdot\|_p$ the p-th vector norm for $p \geq 1$; and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the vector Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

2. PRELIMINARIES AND PROBLEM STATEMENT

Consider the *under-determined* linear system of equations, $\mathbf{y} = \mathbf{A}\mathbf{x}_o$, where the unknown $n \times 1$ vector $\mathbf{x}_o$ is to be recovered from the given $m \times 1$ data vector $\mathbf{y}$, and the $m \times n$ matrix $\mathbf{A}$. If the unknown vector is sparse with many zeros at unknown entries, CS theory ensures quantifiable chances of recovering $\mathbf{x}_o$ even when $m < n$, and the available $\mathbf{y}$ is perturbed [1]. Specifically, the basis pursuit (BP) scheme [2] and its Lagrangian counterpart (namely Lasso) [4,7], both account for the said perturbations. For uniformity, the BP/Lasso solvers can be equivalently written in the form of the least-squares (LS) criterion regularized by the ℓ_1 norm, as

$$\{\hat{\mathbf{x}}, \hat{\mathbf{e}}\}_{Lasso} := \arg\min_{\mathbf{x},\mathbf{e}} \ \|\mathbf{e}\|_2^2 + \lambda_1 \|\mathbf{x}\|_1 \tag{1a}$$

$$\text{s. to } \ \mathbf{y} + \mathbf{e} = \mathbf{A}\mathbf{x}\,. \tag{1b}$$

This work is supported by the NSF grants CCF-0830480, CCF-1016605, ECCS-0824007, and ECCS-1002180; G. Leus is supported in part by NWO-STW under the VICI program (project 10382).

In the context of CS, perturbations in $\mathbf{A}$ can emerge due to disturbances in the compressing measurement matrix, the mismatch in the adopted sparsity expansion basis, or, in both [9]. To account for such perturbations, the S-TLS approach amounts to finding

$$\left\{\hat{\mathbf{x}}, \hat{\mathbf{E}}, \hat{\mathbf{e}}\right\}_{S-TLS} := \arg\min_{\mathbf{x},\mathbf{e},\mathbf{E}} \ \|[\mathbf{E} \ \ \mathbf{e}]\|_F^2 + \lambda\|\mathbf{x}\|_1 \tag{2a}$$

$$\text{s. to } \ \mathbf{y} + \mathbf{e} = (\mathbf{A} + \mathbf{E})\mathbf{x} \tag{2b}$$

where $\lambda > 0$ is a sparsity-tuning constant. Compared to the Lasso in (1), the S-TLS constraint (2b) corrects both $\mathbf{y}$ and $\mathbf{A}$ "parsimoniously," in the sense that the resultant linear system yields a solution with minimal ℓ_1 norm. Similar to LS and BP/Lasso, the S-TLS estimates in (2) are also universal, meaning that perturbations can be random or deterministic, with or without a priori known structure.

However, it is expected that exploiting prior knowledge on the perturbations can only lead to improved performance. Thinking for instance along the lines of weighted LS, one is motivated to *weight* $\|\mathbf{E}\|_F^2$ and $\|\mathbf{e}\|_2^2$ in (2) by their inverse covariance matrices, respectively, whenever those are known and are not both equal to $\mathbf{I}$. As a second motivating example, normal equations, involved in e.g., linear prediction, entail *structure* in $\mathbf{E}$ and/or $\mathbf{e}$ that capture sample estimation errors present in the matrix $[\mathbf{A} \ \ \mathbf{y}]$, which is Toeplitz. Prompted by these examples, the present paper broadens the scope of S-TLS with weighted and structured forms capitalizing on prior knowledge available about $[\mathbf{E} \ \ \mathbf{e}]$, following the spirit of generalizing the TLS to its weighted and structured counterparts in [6]. To this end, it is first prudent to quantify the notion of structure.

Definition 1 *The $m \times (n+1)$ data matrix $[\mathbf{A} \ \ \mathbf{y}](\mathbf{p})$ has structure characterized by an $n_p \times 1$ parameter vector $\mathbf{p}$, if and only if there is a mapping such that $\mathbf{p} \in \mathbb{R}^{n_p} \to [\mathbf{A} \ \ \mathbf{y}](\mathbf{p}) := \mathbf{S}(\mathbf{p}) \in \mathbb{R}^{m\times(n+1)}$.*

Definition 1 is general enough to encompass any (even unstructured) matrix $[\mathbf{A} \ \ \mathbf{y}](\mathbf{p})$, by simply letting $\mathbf{p} := \text{vec}([\mathbf{A} \ \ \mathbf{y}])$ of length $n_p = m(n+1)$. However, it becomes more relevant when $n_p \ll m(n+1)$, the case in which $\mathbf{p}$ characterizes $[\mathbf{A} \ \ \mathbf{y}]$ parsimoniously. Application examples are abundant: structure in Toeplitz and Hankel matrices encountered with system identification, deconvolution, and linear prediction; as well as in circulant and Vandermonde matrices showing up in spatio-temporal harmonic retrieval problems [6]. Structured matrices $\mathbf{A}$ for sparse vectors $\mathbf{x}_o$ emerge also in contemporary CS gridding-based applications e.g., for spectral analysis and estimation of time-varying channels, where rows of the FFT matrix are selected at random; details on other interesting gridding-based cases can be found in [9].

Consider now re-casting the S-TLS criterion in terms of $\mathbf{p}$, and its associated perturbation vector denoted by $\boldsymbol{\epsilon} \in \mathbb{R}^{n_p}$. The Frobenius norm in the cost of (2a) is mapped to the ℓ_2-norm of $\boldsymbol{\epsilon}$; and to allow for weighting the perturbation vector using a symmetric positive definite matrix $\mathbf{W} \in \mathbb{R}^{n_p\times n_p}$, the weighted counterpart of $\|[\mathbf{E} \ \ \mathbf{e}]\|_F^2$ becomes $\boldsymbol{\epsilon}^T\mathbf{W}\boldsymbol{\epsilon}$. With regards to the constraint, Definition 1 implies a perturbed matrix of the form $\mathbf{S}(\mathbf{p}+\boldsymbol{\epsilon}) = [\mathbf{A}+\mathbf{E} \ \ \mathbf{y}+\mathbf{e}]$; hence, re-writing (2b) as $[\mathbf{A}+\mathbf{E} \ \ \mathbf{y}+\mathbf{e}] \left[\mathbf{x}^T, -1\right]^T = \mathbf{0}_m$, yields the structured constraint $\mathbf{S}(\mathbf{p}+\boldsymbol{\epsilon}) \left[\mathbf{x}^T, -1\right]^T = \mathbf{0}_m$. Putting things together, leads to the weighted and structured (WS)S-TLS version of (2)

$$\{\hat{\mathbf{x}}, \hat{\boldsymbol{\epsilon}}\}_{WSS-TLS} := \arg\min_{\mathbf{x},\boldsymbol{\epsilon}} \ \boldsymbol{\epsilon}^T\mathbf{W}\boldsymbol{\epsilon} + \lambda\|\mathbf{x}\|_1 \tag{3a}$$

$$\text{s. to } \ \mathbf{S}(\mathbf{p}+\boldsymbol{\epsilon}) \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \mathbf{0}_m \tag{3b}$$

which clearly subsumes the structured-only form as a special case corresponding to $\mathbf{W} = \mathbf{I}$. With the WSS-TLS problem (3) in mind, the main goal now is to develop efficient algorithms to solve it – a challenging task since presence of the product between $\boldsymbol{\epsilon}$ and $\mathbf{x}$ in (3b) reveals that the problem is generally nonconvex. This consideration motivates focusing on one practically interesting subset of structure mappings. The same subset turns out to simplify the WSS-TLS problem, and subsequently allows asserting MAP optimality for EIV models, and developing efficient solvers based on block coordinate descent.

3. AFFINE AND SEPARABLE STRUCTURES

To confine the structure quantified in Definition 1, two conditions will be imposed. They are also adopted by TLS approaches [6], and are satisfied by the CS applications in [9].

(as1) *The structure mapping in Definition 1 is separable, meaning that with $\mathbf{p} = [(\mathbf{p}^A)^T \ (\mathbf{p}^y)^T]^T$, where $\mathbf{p}^A \in \mathbb{R}^{n_A}$ and $\mathbf{p}^y \in \mathbb{R}^{n_y}$, it holds that $\mathbf{S}(\mathbf{p}) := [\mathbf{A} \ \ \mathbf{y}](\mathbf{p}) = [\mathbf{A}(\mathbf{p}^A) \ \ \mathbf{y}(\mathbf{p}^y)]$. In addition, the separable mapping is linear (more precisely affine), if and only if the $\mathbf{S}(\mathbf{p})$ matrix is composed of known structural elements, namely "matrix atoms" $\{\mathbf{S}_k^A\}_{k=1}^{n_A}$ and "vector atoms" $\{\mathbf{s}_k^y\}_{k=1}^{n_y}$, so that*

$$\mathbf{S}(\mathbf{p}+\boldsymbol{\epsilon}) = \mathbf{S}(\mathbf{p}) + \left[\sum_{k=1}^{n_A} \epsilon_k^A\mathbf{S}_k^A \quad \sum_{k=1}^{n_y} \epsilon_k^y\mathbf{s}_k^y\right] \tag{4}$$

where ϵ_k^A (ϵ_k^y) denotes the k-th entry of the perturbation $\boldsymbol{\epsilon}^A$ ($\boldsymbol{\epsilon}^y$). In accordance to separable structures, the weight matrix $\mathbf{W}$ takes the block diagonal form $\mathbf{W} := bdiag(\mathbf{W}^A, \ \mathbf{W}^y)$, which prevents cross-term costs involving $\boldsymbol{\epsilon}^A$ and $\boldsymbol{\epsilon}^y$ in (3a).

As in Definition 1, interesting structures in (4) are those with $n_A \ll mn$ and/or $n_y \ll m$. (Consider for instance a circulant $m\times n$ matrix $\mathbf{A}$, which can be represented as in (4) using $n_A = m$ matrix atoms.) The separation of entries of $\mathbf{p}$ according to their relations to the data $\mathbf{A}$ or $\mathbf{y}$ decouples $\boldsymbol{\epsilon}^A$ and $\boldsymbol{\epsilon}^y$, and renders $\mathbf{W}$ block diagonal.

The separability and linearity in (as1) simplifies the constraint in (3b) for the given matrix atoms and vector atoms collected for notational brevity in the matrices

$$\boldsymbol{S}^A := [\mathbf{S}_1^A \ \cdots \ \mathbf{S}_{n_A}^A] \ \text{ and } \ \boldsymbol{S}^y := [\mathbf{s}_1^y \ \cdots \ \mathbf{s}_{n_y}^y]. \tag{5}$$

Indeed, the structure in (4) allows one to write the constraint (3b) as: $-\mathbf{S}(\mathbf{p})[\mathbf{x}^T, -1]^T = \mathbf{y} - \mathbf{A}\mathbf{x} = [\sum_{k=1}^{n_A} \epsilon_k^A\mathbf{S}_k^A \ \sum_{k=1}^{n_y} \epsilon_k^y\mathbf{s}_k^y][\mathbf{x}^T, \ -1]^T$; while the latter is compactly denoted by $\boldsymbol{S}^A(\mathbf{I}\otimes\mathbf{x})\boldsymbol{\epsilon}^A - \boldsymbol{S}^y\boldsymbol{\epsilon}^y$, using the definitions in (5) along with the identity $\left(\sum_{k=1}^{n_A} \epsilon_k^A\mathbf{S}_k^A\right)\mathbf{x} = \boldsymbol{S}^A(\boldsymbol{\epsilon}^A \otimes \mathbf{I})\mathbf{x} = \boldsymbol{S}^A(\mathbf{I}\otimes\mathbf{x})\boldsymbol{\epsilon}^A$. In a nutshell, (3b) under (as2) becomes $\boldsymbol{S}^A(\mathbf{I}\otimes\mathbf{x})\boldsymbol{\epsilon}^A - \boldsymbol{S}^y\boldsymbol{\epsilon}^y = \mathbf{y} - \mathbf{A}\mathbf{x}$, in which $\boldsymbol{\epsilon}^A$ is decoupled from $\boldsymbol{\epsilon}^y$. Further, consider the Cholesky decomposition for the two symmetric positive definite weight matrices

$$\mathbf{W}^A = (\boldsymbol{\Gamma}^A)^{-T}(\boldsymbol{\Gamma}^A)^{-1}, \ \text{ and } \ \mathbf{W}^y = (\boldsymbol{\Gamma}^y)^{-T}(\boldsymbol{\Gamma}^y)^{-1}. \tag{6}$$

Correspondingly, the cost in (3a) becomes an unweighted quadratic function for the *normalized* perturbations

$$\boldsymbol{\varepsilon}^A := (\boldsymbol{\Gamma}^A)^{-1}\boldsymbol{\epsilon}^A \ , \ \text{ and } \ \boldsymbol{\varepsilon}^A := (\boldsymbol{\Gamma}^y)^{-1}\boldsymbol{\epsilon}^y \ . \tag{7}$$

Therefore, the WSS-TLS in (3) takes the structured-only form

$$\min_{\mathbf{x},\boldsymbol{\varepsilon}^A,\boldsymbol{\varepsilon}^y} \ \|\boldsymbol{\varepsilon}^A\|_2^2 + \|\boldsymbol{\varepsilon}^y\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{8a}$$

$$\text{s. to } \ \boldsymbol{S}^A(\mathbf{I}\otimes\mathbf{x})\boldsymbol{\Gamma}^A\boldsymbol{\varepsilon}^A - \boldsymbol{S}^y\boldsymbol{\Gamma}^y\boldsymbol{\varepsilon}^y = \mathbf{y} - \mathbf{A}\mathbf{x} \tag{8b}$$

or in a more compact form as: $\min_{\mathbf{x},\boldsymbol{\varepsilon}} \{\|\boldsymbol{\varepsilon}\|_2^2 + \lambda\|\mathbf{x}\|_1\}$ s.to $\boldsymbol{G}(\mathbf{x})\boldsymbol{\varepsilon} = \mathbf{r}(\mathbf{x})$, after defining

$$\boldsymbol{G}(\mathbf{x}) := [\boldsymbol{S}^A(\mathbf{I}\otimes\mathbf{x})\boldsymbol{\Gamma}^A \;\; \boldsymbol{S}^y\boldsymbol{\Gamma}^y] \quad \text{and} \quad \mathbf{r}(\mathbf{x}) := \mathbf{y} - \mathbf{A}\mathbf{x}\,. \quad (9)$$

Interestingly, by eliminating one or two sets of variables in (8), it is possible to establish statistical optimality for a structured EIV system model, and obtain efficient, provably convergent solvers. Those reformulations are given in the following lemma.

Lemma 1: *The constrained WSS-TLS form in* (3) *is equivalent to two unconstrained nonconvex optimization problems:*
(a) one involving $\mathbf{x}$ *and* $\boldsymbol{\varepsilon}^A$*, namely*

$$\{\hat{\mathbf{x}},\hat{\boldsymbol{\varepsilon}}^A\}_{WSS-TLS} = \arg\min_{\mathbf{x},\boldsymbol{\varepsilon}^A} \;\|\boldsymbol{\varepsilon}^A\|_2^2 + \left\|(\boldsymbol{S}^y\boldsymbol{\Gamma}^y)^\dagger[\boldsymbol{S}^A(\mathbf{I}\otimes\mathbf{x})\boldsymbol{\Gamma}^A\boldsymbol{\varepsilon}^A - \mathbf{r}(\mathbf{x})]\right\|_2^2 + \lambda\|\mathbf{x}\|_1\,; \quad (10)$$

and (b) one involving only the variable $\mathbf{x}$*, expressed using* (9)*, as*

$$\hat{\mathbf{x}}_{WSS-TLS} = \arg\min_{\mathbf{x}} \mathbf{r}^T(\mathbf{x})\left[\boldsymbol{G}(\mathbf{x})\boldsymbol{G}^T(\mathbf{x})\right]^\dagger\mathbf{r}(\mathbf{x}) + \lambda\|\mathbf{x}\|_1\,. \quad (11)$$

Both reformulations follow by optimizing over a subset of (inner) variables ($\boldsymbol{\varepsilon}^y$ only in (a), or both $\boldsymbol{\varepsilon}^A$ and $\boldsymbol{\varepsilon}^y$ in (b)), while fixing the rest of the variables. The proof relies on the existence of closed-form solutions for the two aforementioned inner problems, which can be substituted back to (8) to yield the unconstrained formulations in Lemma 1. It will be shown next that the equivalent unconstrained WSS-TLS problem in (10) leads to a MAP optimal estimator which is also obtained efficiently.

4. WSS-TLS OPTIMALITY AND SOLVERS

4.1. MAP Optimality for EIV Models

Consider the structured EIV model with perturbed input ($\mathbf{A}$) and perturbed output ($\mathbf{y}$) obeying the relationship

$$\mathbf{y} = \mathbf{A}(\mathbf{p}_o^A)\mathbf{x}_o + (-\boldsymbol{S}^y\boldsymbol{\epsilon}_y)\,,$$
$$\mathbf{A} = \mathbf{A}(\mathbf{p}_o^A) + \left[-\boldsymbol{S}^A(\boldsymbol{\epsilon}_A\otimes\mathbf{I})\right] \quad (12)$$

where the notation of the model perturbations $\boldsymbol{\epsilon}_A$ and $\boldsymbol{\epsilon}_y$ stresses their difference with the optimization variables $\boldsymbol{\epsilon}^A$ and $\boldsymbol{\epsilon}^y$ in (8). Unknown are the vector $\mathbf{x}_o$, and the inaccessible input matrix $\mathbf{A}(\mathbf{p}_o^A)$, characterized by the vector $\mathbf{p}_o^A$. The WSS-TLS estimator will turn out to be MAP optimal under the following assumption.

(as2) *Perturbations in* (12) *are jointly Gaussian, i.e.,* $[\boldsymbol{\epsilon}_A\;\boldsymbol{\epsilon}_y] \sim \mathcal{N}(\mathbf{0}_{n_p}, \mathbf{W}^{-1})$*, as well as independent from* $\mathbf{p}_o^A$ *and* $\mathbf{x}_o$*. Entries of* $\mathbf{x}_o$ *are zero-mean, i.i.d., according to a common Laplace distribution with common parameter* $2/\lambda$*, and are independent from* $\mathbf{p}_o^A$*, which has i.i.d. entries drawn from a zero-mean uniform (i.e., non-informative) prior pdf.*

Note that the heavy-tailed Laplacian prior on $\mathbf{x}_o$ under (as2) is in par with its "non-probabilistic" sparsity attribute. It has been used to establish that the Lasso estimator $\hat{\mathbf{x}}_{Lasso}$ in (1) is MAP optimal when $\boldsymbol{\epsilon}_A \equiv \mathbf{0}_{n_A}$ and $\boldsymbol{\epsilon}_y$ is white Gaussian with $\boldsymbol{S}^y = \mathbf{I}$ [7]. Moreover, viewing $\mathbf{x}_o$ and $\mathbf{p}_o^A$ as deterministic, the weighted TLS estimator is known to be optimum in the maximum likelihood (ML) sense for the unstructured counterpart of the EIV model in (12) where $\mathbf{p} = \text{vec}([\mathbf{A}\;\;\mathbf{y}])$ [6].

The following optimality claim holds for the WSS-TLS estimator in (10), assured to be equivalent to the solution of problem (8) by Lemma 1. The proof is based on obtaining the log-likelihood and log-prior probability for $\mathbf{x}_o$ and $\mathbf{p}_o^A$, as detailed in [9].

Proposition 1: *(MAP optimality of WSS-TLS). Under (as1) and (as2) and assuming* $\boldsymbol{S}^y$ *of full column rank, the equivalent WSS-TLS problem in* (10) *yields the MAP optimal estimator of* $\mathbf{x}_o$ *and* $\mathbf{p}_o^A$ *in the structured EIV model* (12)*.*

4.2. WSS-TLS Solvers

The formulation in (10) also suggests directly an iterative WSS-TLS solver based on the block coordinate descent method, which alternately optimizes over the variables $\boldsymbol{\varepsilon}^A$ and $\mathbf{x}$. Specifically, suppose that the estimate $\boldsymbol{\varepsilon}^A(i)$ of $\boldsymbol{\varepsilon}^A$ is available at iteration i. Substituting $\boldsymbol{\varepsilon}^A(i)$ into (10) and defining the un-normalized perturbation iterate $\boldsymbol{\epsilon}^A(i) := \boldsymbol{\Gamma}^A\boldsymbol{\varepsilon}^A(i)$ [cf. (7)], allows to estimate $\mathbf{x}$ as

$$\mathbf{x}(i) = \arg\min_{\mathbf{x}} \left\|(\boldsymbol{S}^y\boldsymbol{\Gamma}^y)^\dagger\left[\boldsymbol{S}^A(\boldsymbol{\epsilon}^A(i)\otimes\mathbf{I})\mathbf{x} - \mathbf{r}(\mathbf{x})\right]\right\|_2^2 + \lambda\|\mathbf{x}\|_1\,. \quad (13)$$

Since $\mathbf{r}(\mathbf{x})$ is linear in $\mathbf{x}$ [cf. (9)], the cost in (13) is convex (quadratic regularized by the ℓ_1-norm as in the Lasso cost in (1)); thus, it can be solved efficiently. Likewise, given $\mathbf{x}(i)$ the perturbation vector for the ensuing iteration can be found in closed form since the pertinent cost is quadratic; that is,

$$\boldsymbol{\varepsilon}^A(i+1) = \arg\min_{\boldsymbol{\varepsilon}^A} \;\|\boldsymbol{\varepsilon}^A\|_2^2 + \left\|\check{\boldsymbol{S}}(i)\boldsymbol{\varepsilon}^A - (\boldsymbol{S}^y\boldsymbol{\Gamma}^y)^\dagger\mathbf{r}(\mathbf{x}(i))]\right\|_2^2 \quad (14)$$

where $\check{\boldsymbol{S}}(i) := (\boldsymbol{S}^y\boldsymbol{\Gamma}^y)^\dagger\boldsymbol{S}^A(\mathbf{I}\otimes\mathbf{x}(i))\boldsymbol{\Gamma}^A$. To express $\boldsymbol{\varepsilon}^A(i+1)$ compactly, equating to zero the gradient of the cost in (14), yields

$$\boldsymbol{\varepsilon}^A(i+1) = \left[\mathbf{I} + \check{\boldsymbol{S}}^T(i)\check{\boldsymbol{S}}(i)\right]^{-1}\check{\boldsymbol{S}}^T(i)(\boldsymbol{S}^y\boldsymbol{\Gamma}^y)^\dagger\mathbf{r}(\mathbf{x}(i))\,. \quad (15)$$

Initialized with $\boldsymbol{\varepsilon}^A(0) = \mathbf{0}_{n_A}$, the algorithm cycles between iterations (13) and (15). Using the basic convergence result in [8], it can be shown that these iterations are convergent as asserted in the following; see [9, Prop. 3] for detailed arguments.

Proposition 2: *(Convergence). Under (as1), the iterates in* (13) *and* (15) *converge monotonically at least to a stationary point of the unconstrained WSS-TLS problem in* (10)*.*

As estimating $\boldsymbol{\varepsilon}^A$ is simple using the closed form in (15), it is useful at this point to explore tailored solvers for the Lasso-type problem in (13). The coordinate descent (CD) is a popular choice for this purpose; see e.g., [4]. In the present context, CD cycles between $\boldsymbol{\varepsilon}^A(i)$, and scalar iterates of the $\mathbf{x}(i)$ entries. To update the ν-th entry $x_\nu(i)$, suppose precursor entries $\{x_1(i),\ldots,x_{\nu-1}(i)\}$ have been already obtained in the i-th iteration, and postcursor entries $\{x_{\nu+1}(i-1),\ldots,x_n(i-1)\}$ are also available from the previous $(i-1)$-st iteration along with $\boldsymbol{\varepsilon}^A(i)$, found in closed form as in (15). Letting $\boldsymbol{\alpha}_\nu(i)$ denote the ν-th column of $\left[(\boldsymbol{S}^y\boldsymbol{\Gamma}^y)^\dagger\left(\mathbf{A} + \boldsymbol{S}^A(\boldsymbol{\epsilon}^A(i)\otimes\mathbf{I})\right)\right]$, the effect of these known entries can be removed from $\mathbf{y}$ by forming

$$\mathbf{e}_\nu(i) := (\boldsymbol{S}^y\boldsymbol{\Gamma}^y)^\dagger\mathbf{y} - \textstyle\sum_{j=1}^{\nu-1}\boldsymbol{\alpha}_j(i)x_j(i) - \sum_{j=\nu+1}^{n}\boldsymbol{\alpha}_j(i)x_j(i-1)\,. \quad (16)$$

Using (16), the vector optimization in (13) now reduces to the following scalar one: $x_\nu(i) = \arg\min_{x_\nu}\{\|\boldsymbol{\alpha}_\nu(i)x_\nu - \mathbf{e}_\nu(i)\|_2^2 +$

$\lambda|x_\nu|\}$. This *scalar* Lasso problem is known to admit a closed-form solution expressed in terms of a soft-thresholding operator

$$x_\nu(i) = \text{sign}\left(\mathbf{e}_\nu^T(i)\boldsymbol{\alpha}_\nu(i)\right)\left[\frac{\left|\mathbf{e}_\nu^T(i)\boldsymbol{\alpha}_\nu(i)\right|}{\|\boldsymbol{\alpha}_\nu(i)\|_2^2} - \frac{\lambda}{2\|\boldsymbol{\alpha}_\nu(i)\|_2^2}\right]_+ \quad (17)$$

where $\text{sign}(\cdot)$ denotes the sign operator, and $[\chi]_+ := \chi$, if $\chi > 0$, and zero otherwise.

Cycling through the closed forms (15)-(17) explains why CD here is faster than, and thus preferable over general-purpose convex optimization solvers of (13). Another factor contributing to its speed is the sparsity of $\mathbf{x}(i)$, which implies that starting up with the all-zero vector, namely $\mathbf{x}(-1) = \mathbf{0}_n$, offers initialization close to a stationary point of the cost in (10). Convergence to this stationary point is guaranteed by using the results in [8], similar to Proposition 2. Note also that larger values of λ in (17) force more entries of $\mathbf{x}(i)$ to be shrunk to zero, which corroborates the role of λ as a sparsity-tuning parameter.

The CD based WSS-TLS solver is tabulated as Algorithm 1.

Algorithm 1 : CD for WSS-TLS

Initialize with $\boldsymbol{\varepsilon}(0) = \mathbf{0}_{n_A}$ and $\mathbf{x}(-1) = \mathbf{0}_n$
for $i = 0, 1, \ldots$ **do**
  for $\nu = 1, \ldots, n$ **do**
    Compute the residual $\mathbf{e}_\nu(i)$ as in (16).
    Update the scalar $x_\nu(i)$ via (17).
  end for
  Update the iterate $\boldsymbol{\varepsilon}^A(i+1)$ as in (15).
end for

5. NUMERICAL EXAMPLES

To illustrate the merits of the WSS-TLS algorithm, consider a deconvolution problem setup, where the 20×40 regression matrix $\mathbf{A}$ has non-symmetric Toeplitz structure [6]. A structured EIV model in (12) is randomly generated per trial, with the $20 + 40 - 1 = 59$ entries of $\mathbf{p}_o^A$ independent, zero-mean, Gaussian distributed with variance $1/20$ (so that on average each column of $\mathbf{A}$ has unit ℓ_2 norm). Vector $\mathbf{x}_o$ has only 10 nonzero entries drawn from the standardized Gaussian distribution. Following (as2), $\boldsymbol{\epsilon}_A$ and $\boldsymbol{\epsilon}_y$ in (12) are independent white Gaussian vectors with corresponding entry-wise variances $(0.15)^2/20$ and $(0.05)^2/20$. The Lasso and various (WS)S-TLS estimates are compared over 100 trials and 20 values of λ uniformly distributed in log-scale. The structured S-TLS only accounts for the Toeplitz structure with $\mathbf{W} = \mathbf{I}$ in (8), while the weighted one refers to the unstructured WSS-TLS with $\mathbf{p} = \text{vec}([\mathbf{A} \;\; \mathbf{y}])$. Fig. 1 compares their empirical ℓ_0 and ℓ_2 errors [1] in estimating $\mathbf{x}_o$.

Fig. 1 illustrates that all (W/S)S-TLS approaches outperform the Lasso one, especially in recovering the support based on the ℓ_0 error. As shown in Fig. 1(a), the two structured S-TLS variants improve over their unstructured versions, while WSS-TLS achieves smaller error compared to the structured-only S-TLS. Nevertheless, as λ increases all estimates degrade equally. This is because the increasing sparsity penalty favors the all-zero solution. Although the Lasso estimate exhibits smaller errors at the tail, the S-TLS variants, especially the WSS-TLS one still shows a slight edge over the range of moderate λ values. More practical examples and tests are available in [9], which shows applicability of (WS)S-TLS in calibrating the mismatch effects of contemporary grid-based approaches to cognitive radio sensing, and high-resolution direction-of-arrival estimation using antenna arrays.

[1] The ℓ_0 error records the percentage of entries where the support of the two vectors differs; the ℓ_1 error plot is similar to the ℓ_2 one in comparing these methods, and is omitted due to space limitation.

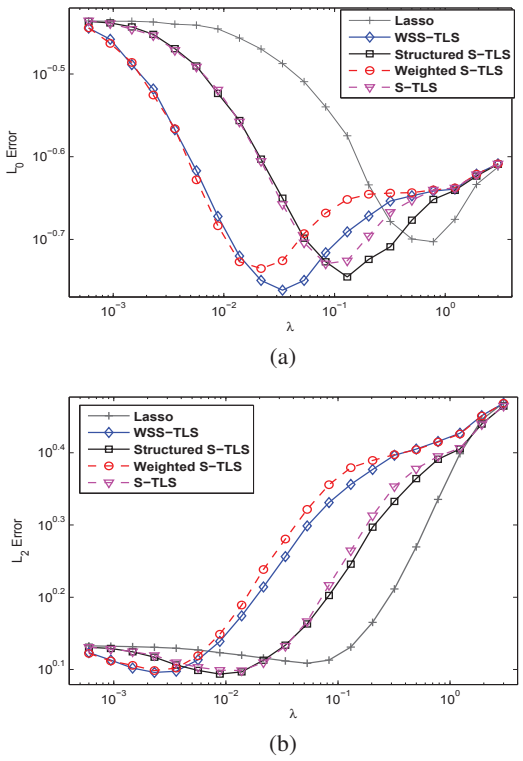

Fig. 1. Comparisons among the Lasso and variants of (WS)S-TLS estimates for the (a) ℓ_0 error; and (b) ℓ_2 error.

6. REFERENCES

[1] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, Jul. 2007.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. on Sci. Comp.*, vol. 20, pp. 33–61, Jan. 1998.

[3] Y. Chi, A. Pezeshki, L. Scharf, and R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," in *Proc. of ICASSP*, Mar. 2010.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed., 2009.

[5] M. A. Herman and T. Strohmer, "General deviants: an analysis of perturbations in compressive sensing," *IEEE J. of Selected Topics in Signal Processing*, vol. 4, pp. 342–349, Apr. 2010.

[6] I. Markovsky and S. Van Huffel, "Overview of total least-squares methods," *Signal Processing*, vol. 87, no. 10, pp. 2283–2302, Oct. 2007.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc. B*, vol. 58, pp. 267–288, 1996.

[8] P. Tseng, "Convergence of a block coordinate descent method for non-differentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, Jun. 2001.

[9] H. Zhu, G. Leus, and G. B. Giannakis, "Sparsity-cognizant total least-squares for perturbed compressive sampling," *IEEE TSP*, Aug. 2010, (submitted). [Online]. Available: http://arxiv.org/abs/1008.2996