6-2010

Variance-Component Based Sparse Signal Reconstruction and Model Selection

Kun Qiu
Iowa State University

Aleksandar Dogandžić
Iowa State University, ald@iastate.edu

Variance-Component Based Sparse Signal Reconstruction and Model Selection

Abstract
We propose a variance-component probabilistic model for sparse signal reconstruction and model selection. The measurements follow an underdetermined linear model, where the unknown regression vector (signal) is sparse or approximately sparse and noise covariance matrix is known up to a constant. The signal is composed of two disjoint parts: a part with significant signal elements and the complementary part with insignificant signal elements that have zero or small values. We assign distinct variance components to the candidates for the significant signal elements and a single variance component to the rest of the signal; consequently, the dimension of our model's parameter space is proportional to the assumed sparsity level of the signal. We derive a generalized maximum-likelihood (GML) rule for selecting the most efficient parameter assignment and signal representation that strikes a balance between the accuracy of data fit and compactness of the parameterization. We prove that, under mild conditions, the GML-optimal index set of the distinct variance components coincides with the support set of the sparsest solution to the underlying underdetermined linear system. Finally, we propose an expansion-compression variance-component based method (ExCoV) that aims at maximizing the GML objective function and provides an approximate GML estimate of the significant signal element set and an empirical Bayesian signal estimate. The ExCoV method is automatic and demands no prior knowledge about signal-sparsity or measurement-noise levels. We also develop a computationally and memory efficient approximate ExCoV scheme suitable for large-scale problems, apply the proposed methods to reconstruct one- and two-dimensional signals from compressive samples, and demonstrate their reconstruction performance via numerical simulations. Compared with the competing approaches, our schemes perform particularly well in challenging scenarios where the noise is large or the number of measurements is small.

Keywords
Bayes methods, Covariance matrices, Data compression, Maximum likelihood estimation, Probability

Disciplines
Signal Processing

Comments
This is a post-print of an article from *IEEE Transactions on Signal Processing* 58 (2010): 2935–2952, doi:10.1109/TSP.2010.2044828. Posted with permission.

Variance-Component Based Sparse Signal Reconstruction and Model Selection

Kun Qiu and *Aleksandar Dogandžić*
ECpE Department, Iowa State University, 3119 Coover Hall, Ames, IA 50011
Phone: (515) 294-0500, Fax: (515) 294-8432, email: {kqiu,ald}@iastate.edu

Abstract

We propose a variance-component probabilistic model for sparse signal reconstruction and model selection. The measurements follow an underdetermined linear model, where the unknown regression vector (signal) is sparse or approximately sparse and noise covariance matrix is known up to a constant. The signal is composed of two disjoint parts: a part with significant signal elements and the complementary part with insignificant signal elements that have zero or small values. We assign distinct variance components to the candidates for the significant signal elements and a single variance component to the rest of the signal; consequently, the dimension of our model's parameter space is proportional to the assumed sparsity level of the signal. We derive a generalized maximum likelihood (GML) rule for selecting the most efficient parameter assignment and signal representation that strikes a balance between the accuracy of data fit and compactness of the parameterization. We prove that, under mild conditions, the GML-optimal index set of the distinct variance components *coincides* with the support set of the sparsest solution to the underlying underdetermined linear system. Finally, we propose an expansion-compression variance-component based method (ExCoV) that aims at maximizing the GML objective function and provides an approximate GML estimate of the significant signal element set and an empirical Bayesian signal estimate. The ExCoV method is *automatic* and demands no prior knowledge about signal-sparsity or measurement-noise levels. We also develop a computationally and memory efficient approximate ExCoV scheme suitable for large-scale problems, apply the proposed methods to reconstruct one- and two-dimensional signals from compressive samples, and demonstrate their reconstruction performance via numerical simulations. Compared with the competing approaches, our schemes perform particularly well in challenging scenarios where the noise is large or the number of measurements is small.

I. Introduction

Over the past decade, sparse signal processing methods have been developed and successfully applied to biomagnetic imaging, spectral and direction-of-arrival estimation, and compressive sampling, see [1]–[7] and references therein. Compressive sampling is an emerging signal acquisition and processing paradigm that allows perfect reconstruction of sparse signals from highly undersampled measurements. Compressive sampling and sparse-signal reconstruction will likely play a pivotal role in accommodating the rapidly expanding digital data space.

For noiseless measurements, the major sparse signal reconstruction task is finding the sparsest solution of an underdetermined linear system $\boldsymbol{y} = H\,\boldsymbol{s}$ (see e.g. [7, eq. (2)]):

$$(\mathrm{P}_0): \qquad \min_{\boldsymbol{s}} \|\boldsymbol{s}\|_{\ell_0} \quad \text{subject to } \boldsymbol{y} = H\,\boldsymbol{s} \tag{1.1}$$

where $\boldsymbol{y}$ is an $N \times 1$ measurement vector, $\boldsymbol{s}$ is an $m \times 1$ vector of unknown *signal coefficients*, H is a known $N \times m$ full-rank *sensing matrix* with $N < m$, and $\|\boldsymbol{s}\|_{\ell_0}$ counts the number of nonzero elements in the signal vector $\boldsymbol{s}$. The (P_0) problem requires combinatorial search and is known to be NP-hard [8]. Many tractable approaches have been proposed to find sparse solutions to the above underdetermined system. They can be roughly divided into four groups: convex relaxation, greedy pursuit, probabilistic, and other methods.

The main idea of convex relaxation is to replace the ℓ_0-norm penalty with the ℓ_1-norm penalty and solve the resulting convex optimization problem. Basis pursuit (BP) directly substitutes ℓ_0 with ℓ_1 in the (P_0) problem, see [9]. To combat measurement noise and accommodate for approximately sparse signals, several methods with various optimization objectives have been suggested, e.g. basis pursuit denoising (BPDN) [5], [9], Dantzig selector [10], least absolute shrinkage and selection operator (LASSO) [11], and gradient projection for sparse reconstruction (GPSR) [12]. The major advantage of these methods is the uniqueness of their solution due to the convexity of the underlying objective functions. However, this unique global solution generally *does not* coincide with the solution to the (P_0) problem in (1.1): using the ℓ_1-norm penalizes larger signal elements more, whereas the ℓ_0-norm imposes the same penalty on all non-zeros. Moreover, most convex methods require tuning, where the tuning parameters are typically functions of the noise or signal sparsity levels. Setting the tuning parameters is not trivial and the reconstruction performance depends crucially on their choices.

Greedy pursuit methods approximate the (P_0) solution in an iterative manner by making locally optimal choices. An early method from this group is orthogonal matching pursuit (OMP) [13], [14], which adds a single element per iteration to the estimated sparse-signal support set so that a squared-error criterion is minimized. However, OMP achieves limited success in reconstructing sparse signals. To improve the reconstruction performance or complexity of OMP, several OMP variants have been recently developed, e.g. stagewise OMP [15], compressive sampling matching pursuit (CoSaMP) [16], and subspace pursuit [17]. However, greedy methods also require tuning, with tuning parameters related to the signal sparsity level.

Probabilistic methods utilize full probabilistic models. Many popular sparse recovery schemes can be interpreted using a probabilistic point of view. For example, basis pursuit yields the maximum *a posteriori* (MAP) signal estimator under a Bayesian model with sparse-inducing Laplace prior distribution. The most popular probabilistic approaches include sparse Bayesian learning (SBL) [18], [19] and Bayesian compressive sensing (BCS) [20]. SBL adopts an *empirical Bayesian* approach and employs a Gaussian prior on the signal, with a distinct variance component on each signal element; these variance components are estimated by maximizing a marginal likelihood function via the expectation-maximization (EM) algorithm. This marginal likelihood function is globally optimized by variance component estimates that correspond to the (P_0)-optimal signal support, see [18, Theorem 1] and [19, Result 1] and Corollary 5 in Appendix B. Our experience with numerical experiments indicates that SBL achieves the top-tier performance compared with the state-of-art reconstruction methods. Moreover, unlike many other approaches that require tuning, SBL is *automatic* and does not require tuning or knowledge of signal sparsity or noise levels. The major shortcomings of SBL are its high computational complexity and large memory requirements, which make its application on large-scale data (e.g. images and video) practically impossible. SBL needs EM iterations over a parameter space of dimension $m+1$, and most of parameters converge to zero and are redundant. This makes SBL significantly slower

than other sparse signal reconstruction techniques. The BCS method in [20] stems from relevance vector machines [21] and can be understood as a variational formulation of SBL [18, Sec. V]. BCS circumvents the EM iteration and is much faster than SBL, at a cost of poorer reconstruction performance.

We now discuss other methods that cannot be classified into the above three groups. Iterative hard thresholding (IHT) schemes [22]–[25] apply simple iteration steps that do not involve matrix inversions. However, IHT schemes often need good initial values to start the iteration and require tuning, where the signal sparsity level is a typical tuning parameter [24], [26]. Interestingly, the IHT method in [24] can be cast into the probabilistic framework, see [27]. Focal underdetermined system solver (FOCUSS) [1] repeatedly solves a weighted ℓ_2-norm minimization, with larger weights put on the smaller signal components. Although close to the (P_0) problem in the objective function, FOCUSS suffers from abundance of local minima, which limits its reconstruction performance [18], [28]. Analogous to FOCUSS, reweighted ℓ_1 minimization iteratively solves a weighted basis pursuit problem [29], [30]; in [29], this approach is reported to achieve better reconstruction performance than BP, where the runtime of the former is multiple times that of the latter. A sparsity related tuning parameter is also needed to ensure the stability of the reweighted ℓ_1 method [29, Sec. 2.2].

The contribution of this paper is three-fold.

First, we propose a probabilistic model that generalizes the SBL model and, typically, has a much smaller number of parameters than SBL. This generalization makes full use of the key feature of sparse or approximately sparse signals, that most signal elements are zero or close to zero, and only a few have nontrivial magnitudes. Therefore, the signal is naturally partitioned into the *significant* and the complementary *insignificant* signal elements. Rather than allocating individual variance-component parameters to all signal elements, we only assign distinct variance components to the candidates for significant signal elements and a single variance component to the rest of the signal. Consequently, the dimension of our model's parameter space is proportional to the assumed sparsity level of the signal. The proposed model provided a framework for model selection.

Second, we derive a generalized maximum likelihood (GML) rule[1] to select the most efficient parameter assignment under the proposed probabilistic model and prove that, under mild conditions, the GML objective function for the proposed model is globally maximized at the support set of the (P_0) solution. In a nutshell, we have transformed the original constrained (P_0) optimization problem into an equivalent *unconstrained* optimization problem. Unlike the SBL cost function that does not quantify the efficiency of the signal representation, our GML rule evaluates both how compact the signal representation is and how well the corresponding best signal estimate fits the data.

Finally, we propose an expansion-compression variance-component based method (EXCOV) that aims at maximizing

[1]See [31, p. 223 and App. 6F] for general formulation of the GML rule. The GML rule is closely related to *stochastic information complexity*, see [32, eq. (17)] and [33] and references therein.

the GML objective function under the proposed probabilistic model, and provides an empirical Bayesian signal estimate under the selected variance component assignment, see also [34]. In contrast with most existing methods, EXCOV is an automatic algorithm that does not require tuning or knowledge of signal sparsity or noise levels and does not employ a convergence tolerance level or threshold to terminate. Thanks to the parsimony of our probabilistic model, EXCOV is typically significantly faster than SBL, particularly in large-scale problems, see also Section IV-C. We also develop a memory and computationally efficient approximate EXCOV scheme that only involves matrix-vector operations. Various simulation experiments show that, compared with the competing approaches, EXCOV performs particularly well in challenging scenarios where the noise is large or the number of measurements is small, see also the numerical examples in [34].

In Section II, we introduce our variance-component modeling framework and, in Section III, present the corresponding GML rule and our main theoretical result establishing its relationship to the (P_0) problem. In Section IV, we describe the EXCOV algorithm and its efficient approximation (Section IV-B) and contrast their memory and computational requirements with those of the SBL method (Section IV-C). Numerical simulations in Section V compare reconstruction performances of the proposed and existing methods. Concluding remarks are given in Section VI.

A. Notation

We introduce the notation used in this paper:

- $\mathcal{N}(\boldsymbol{y}\,;\,\boldsymbol{\mu},\Sigma)$ denotes the multivariate probability density function (pdf) of a real-valued Gaussian random vector $\boldsymbol{y}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ;
- $|\cdot|$, abs$(\cdot)$, $\|\cdot\|_{\ell_p}$, and "^{T}" denote the determinant, absolute value, ℓ_p norm, and transpose, respectively;
- $\mathrm{card}(A)$ denotes the cardinality of the set A;
- $\lfloor x \rfloor$ is the largest integer smaller than or equal to x;
- I_n and $\mathbf{0}_{n\times 1}$ are the identity matrix of size n and the $n\times 1$ vector of zeros, respectively;
- $\mathrm{diag}\{x_1, x_2, \ldots, x_n\}$ is the $n\times n$ diagonal matrix with the (i,i)th diagonal element x_i, $i=1,2,\ldots,n$;
- "$\odot$" and "$^{\odot 2}$" denote the Hadamard (elementwise) matrix product and elementwise square of a matrix;
- $X^\dagger$ denotes the Moore-Penrose inverse of a matrix X;
- $\Pi(X)$ denotes the projection matrix onto the column space of an $n\times m$ matrix X and $\Pi^\perp(X) = I_n - \Pi(X)$ is the corresponding complementary projection matrix;
- $X \succ Y$ denotes that each element of X is greater than the corresponding element of Y, for equal-size X, Y;
- $[X]_{i,j}$ denotes the (i,j)th element of X;
- $\Sigma^{1/2}$ denotes the Hermitian square root of a covariance matrix Σ and $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$.

II. The Variance-component Probabilistic Measurement Model

We model the pdf of a measurement vector $\boldsymbol{y} \in \mathcal{R}^N$ given $\boldsymbol{s}$ and σ^2 using the standard additive Gaussian noise model:

$$p(\boldsymbol{y} \,|\, \boldsymbol{s}, \sigma^2) = \mathcal{N}\big(\boldsymbol{y}\,;\, H\,\boldsymbol{s}, \sigma^2\, C\big) \tag{2.1}$$

where $H \in \mathcal{R}^{N\times m}$ is the known full-rank *sensing matrix* with

$$N \ll m \tag{2.2}$$

$\boldsymbol{s} \in \mathcal{R}^m$ is an unknown sparse or approximately sparse signal vector, C is a known positive definite symmetric matrix of size $N \times N$, σ^2 is an unknown noise-variance parameter, and $\sigma^2\, C$ is the noise covariance matrix.[2] Setting $C = I_N$ gives white Gaussian noise.

A. Prior Distribution on the Sparse Signal $\boldsymbol{s}$

A prior distribution for the signal $\boldsymbol{s}$ should capture its key feature: sparsity. We know *a priori* that only a few elements of $\boldsymbol{s}$ have nontrivial magnitudes and that the remaining elements are either strictly zero or close to zero. Therefore, $\boldsymbol{s}$ is naturally partitioned into the significant and insignificant signal components. For example, for a strictly sparse signal, its significant part corresponds to the non-zero signal elements, whereas its insignificant part consists of the zero elements, see also Fig. 1 in Section V. The significant signal elements vary widely in magnitude and sign; in contrast, the insignificant signal elements have small magnitudes. We therefore assign *distinct* variance components to the candidates for significant signal elements and use only one common variance-component parameter to account for the variability of the rest of signal coefficients.

Denote by A the set of indices of the signal elements with *distinct* variance components. The set A is *unknown*, with *unknown* size m_A. We also define the complementary index set

$$B = \mathcal{A} \setminus A \tag{2.3a}$$

with cardinality $m_B = m - m_A$, corresponding to signal elements that share a common variance, where

$$\mathcal{A} = \{1, 2, \ldots, m\} \tag{2.3b}$$

denotes the *full index set*. We accordingly partition H and $\boldsymbol{s}$ into submatrices $H_A \in \mathcal{R}^{N\times m_A}$ and $H_B \in \mathcal{R}^{N\times m_B}$, and subvectors $\boldsymbol{s}_A \in \mathcal{R}^{m_A}$ and $\boldsymbol{s}_B \in \mathcal{R}^{m_B}$. Specifically,

- H_A is the *restriction* of the sensing matrix H to the index set A, e.g. if $A = \{1, 2, 5\}$, then $H_A = [\boldsymbol{h}_1\, \boldsymbol{h}_2\, \boldsymbol{h}_5]$, where $\boldsymbol{h}_i$ is the ith column of H and
- $\boldsymbol{s}_A$ is the restriction of the signal-coefficient vector $\boldsymbol{s}$ to the index set A, e.g. if $A = \{1, 2, 5\}$, then $\boldsymbol{s}_A = [s_1, s_2, s_5]^T$, where s_i is the ith element of $\boldsymbol{s}$.

[2] An extension of the proposed approach to circularly-symmetric complex Gaussian measurements, sensing matrix, and signal coefficients is straightforward.

We adopt the following prior model for the signal coefficients:

$$p(\boldsymbol{s}\,|\,\boldsymbol{\delta}_A,\gamma^2) = p(\boldsymbol{s}_A\,|\,\boldsymbol{\delta}_A)\cdot p(\boldsymbol{s}_B\,|\,\gamma^2) = \mathcal{N}\big(\boldsymbol{s}_A\,;\,\mathbf{0}_{m_A\times 1}, D_A(\boldsymbol{\delta}_A)\big)\cdot\mathcal{N}\big(\boldsymbol{s}_B\,;\,\mathbf{0}_{m_B\times 1}, D_B(\gamma^2)\big) \tag{2.4a}$$

where the signal covariance matrices are *diagonal*:

$$D_A(\boldsymbol{\delta}_A) = \mathrm{diag}\{\delta_{A,1}^2,\delta_{A,2}^2,\ldots,\delta_{A,m_A}^2\},\quad D_B(\gamma^2) = \gamma^2\, I_{m_B} \tag{2.4b}$$

with

$$\boldsymbol{\delta}_A = [\delta_{A,1}^2,\delta_{A,2}^2,\ldots,\delta_{A,m_A}^2]^T. \tag{2.4c}$$

The variance components $\delta_{A,1}^2,\delta_{A,2}^2,\ldots,\delta_{A,m_A}^2$ for $\boldsymbol{s}_A$ are *distinct*; the common variance γ^2 accounts for the variability of $\boldsymbol{s}_B$.

The larger A is, the more parameters are introduced to the model. If all signal variance components are freely adjustable, i.e. when $A=\mathcal{A}$ and $m_A=m$, we refer to it as the *full model*. Further, if $C=I_N$, our full model reduces to that of the SBL model in [18] (see also [20, Sec. III] and references therein).

B. Log-likelihood Function of the Variance Components

We assume that the signal variance components $\boldsymbol{\delta}_A$ and γ^2 are *unknown* and define the set of all unknowns:

$$\boldsymbol{\theta} = (A,\boldsymbol{\rho}_A) \tag{2.5a}$$

where

$$\boldsymbol{\rho}_A = (\boldsymbol{\delta}_A,\gamma^2,\sigma^2) \tag{2.5b}$$

is the set of variance-component parameters for a given index set A. The marginal pdf of the observations $\boldsymbol{y}$ given $\boldsymbol{\theta}$ is [see (2.1) and (2.4)]

$$p(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \int p(\boldsymbol{y}\,|\,\boldsymbol{s},\sigma^2)\cdot p(\boldsymbol{s}\,|\,\boldsymbol{\delta}_A,\gamma^2)\,d\boldsymbol{s} = \mathcal{N}\big(\boldsymbol{y}\,;\,\mathbf{0}_{N\times 1},\,P^{-1}(\boldsymbol{\theta})\big) \tag{2.6a}$$

where $P(\boldsymbol{\theta})$ is the precision (inverse covariance) matrix of $\boldsymbol{y}$ given $\boldsymbol{\theta}$:

$$P(\boldsymbol{\theta}) = [H_A\,D_A(\boldsymbol{\delta}_A)\,H_A^T + \gamma^2\,H_B\,H_B^T + \sigma^2\,C]^{-1} \tag{2.6b}$$

and the log-likelihood function of $\boldsymbol{\theta}$ is

$$\ln p(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = -\tfrac{1}{2}\,N\,\ln(2\,\pi) + \tfrac{1}{2}\,\ln|P(\boldsymbol{\theta})| - \tfrac{1}{2}\,\boldsymbol{y}^T\,P(\boldsymbol{\theta})\,\boldsymbol{y}. \tag{2.6c}$$

For a given model A, we can maximize (2.6c) with respect to the model parameters $\boldsymbol{\rho}_A$ using the EM algorithm presented in Section IV-A and derived in Appendix A.

III. GML Rule and Its Equivalence To the (P_0) Problem

We introduce the GML rule for selecting the best index set A, i.e. the best model, see also [31, p. 223]. The best model strikes a balance between fitting the observations $\boldsymbol{y}$ well and keeping the number of model parameters small. The GML rule maximizes

$$\mathrm{GML}(A) = \mathrm{GL}\big((A, \widehat{\boldsymbol{\rho}}_A)\big) \tag{3.1}$$

with respect to A, where

$$\mathrm{GL}(\boldsymbol{\theta}) = \ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) - \tfrac{1}{2} \ln |\mathcal{I}(\boldsymbol{\theta})| \tag{3.2}$$

$\widehat{\boldsymbol{\rho}}_A$ is the ML estimate of $\boldsymbol{\rho}_A$ for *given* A:

$$\widehat{\boldsymbol{\rho}}_A = (\widehat{\boldsymbol{\delta}}_A, \widehat{\gamma}^2, \widehat{\sigma}^2(A)) = \arg\max_{\boldsymbol{\rho}_A} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) \tag{3.3}$$

and $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information matrix (FIM) for the signal variance components $\boldsymbol{\delta}_A$ and γ^2. Since the pdf of $\boldsymbol{y}$ given $\boldsymbol{\theta}$ (2.6a) is Gaussian, we can easily compute $\mathcal{I}(\boldsymbol{\theta})$ using the FIM result for the Gaussian measurement model [35, eq. (3.32) on p. 48]:

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{bmatrix} \mathcal{I}_{\boldsymbol{\delta}_A, \boldsymbol{\delta}_A}(\boldsymbol{\theta}) & \mathcal{I}_{\boldsymbol{\delta}_A, \gamma^2}(\boldsymbol{\theta}) \\ \mathcal{I}^T_{\boldsymbol{\delta}_A, \gamma^2}(\boldsymbol{\theta}) & \mathcal{I}_{\gamma^2, \gamma^2}(\boldsymbol{\theta}) \end{bmatrix} \tag{3.4a}$$

with blocks computed as follows:

$$\mathcal{I}_{\boldsymbol{\delta}_A, \boldsymbol{\delta}_A}(\boldsymbol{\theta}) = \tfrac{1}{2} \left[H_A^T \, P(\boldsymbol{\theta}) \, H_A\right] \odot \left[H_A^T \, P(\boldsymbol{\theta}) \, H_A\right] = \tfrac{1}{2} \left[H_A^T \, P(\boldsymbol{\theta}) \, H_A\right]^{\odot 2} \tag{3.4b}$$

$$[\mathcal{I}_{\boldsymbol{\delta}_A, \gamma^2}(\boldsymbol{\theta})]_i = \tfrac{1}{2} \, H_A^T(:, i) \, P(\boldsymbol{\theta}) \, H_B \, H_B^T \, P(\boldsymbol{\theta}) \, H_A(:, i), \quad i = 1, 2, \ldots, m_A \tag{3.4c}$$

$$\mathcal{I}_{\gamma^2, \gamma^2}(\boldsymbol{\theta}) = \tfrac{1}{2} \, \mathrm{tr}[P(\boldsymbol{\theta}) \, H_B \, H_B^T \, P(\boldsymbol{\theta}) \, H_B \, H_B^T] \tag{3.4d}$$

where $H_A(:, i)$ denotes the ith column of H_A.

The first term in (3.2) is simply the log-likelihood function (2.6c), which evaluates how well the parameters fit the observations. To achieve the best fit for any given model A, we maximize this term with respect to $\boldsymbol{\rho}_A$, see (3.1). The more parameters we have, the better we can fit the measurements. Since any index set A is a subset of the full set $\mathcal{A}$, the maximized log likelihood for $A = \mathcal{A}$ must have larger value than the maximized log likelihood for any other A. However, the second term in (3.2) *penalizes* the growth of A. The GML rule thereby balances modeling accuracy and efficiency.

A. Equivalence of the GML Rule to the (P_0) Problem

We now establish the equivalence between the *unconstrained* GML objective and the *constrained* (P_0) optimization problem (1.1). Let $\boldsymbol{s}^\diamond$ denote the solution to (P_0) and $A^\diamond$ the index set of the nonzero elements of $\boldsymbol{s}^\diamond$, also known as the *support* of $\boldsymbol{s}^\diamond$; then, $m_{A^\diamond} = \|\boldsymbol{s}^\diamond\|_{\ell_0}$ denote the cardinality of $A^\diamond$.

Theorem 1: Assume (2.2) and that

(1) the sensing matrix H satisfies the unique representation property (URP) [1] stating that all $N \times N$ submatrices of H are invertible,

(2) the Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})$ for the signal variance components in (3.4a) is always nonsingular,

(3) the number of measurements N satisfies

$$N > 2\, m_{A^\diamond} + 2. \tag{3.5}$$

Then, the support $A^\diamond$ of the (P_0)-optimal signal-coefficient vector $\boldsymbol{s}^\diamond$ coincides with the GML-optimal index set A, i.e. $\mathrm{GML}(A)$ in (3.1) is *globally and uniquely maximized* at $A = A^\diamond$, and the (P_0)-optimal solution $\boldsymbol{s}^\diamond$ coincides with the empirical Bayesian signal estimate obtained by substituting $A = A^\diamond$ and the corresponding ML variance-component estimates into $\mathrm{E}_{\boldsymbol{s}\,|\,\boldsymbol{y},\boldsymbol{\theta}}[\boldsymbol{s}\,|\,\boldsymbol{y},\boldsymbol{\theta}]$.

Proof: See Appendix B. □

Theorem 1 states that, under conditions (1)–(3), the GML rule is globally maximized at the support set of the (P_0) problem solution; hence, the GML rule *transforms* the constrained optimization problem (P_0) in (1.1) into an equivalent unconstrained optimization problem (3.1). Observe that condition (3) holds when there is no noise and the true underlying signal is sufficiently sparse. Hence, for the noiseless case, Theorem 1 shows that GML-optimal signal model allocates all *distinct* variance components to the nonzero signal elements of the (P_0) solution.

The GML rule allows us to compare signal models. This is not the case for the (P_0) reconstruction approach (1.1) or SBL. The (P_0) approach optimizes a constrained objective and, therefore, does not provide a model evaluation criterion; the SBL objective function, which is the marginal log-likelihood function (2.6c) under the full model $A = \mathcal{A}$ and $C = I_N$, has a fixed number of parameters and, therefore, does not compare different signal models.

In Sections IV, we develop our EXCOV scheme for approximating the GML rule.

IV. THE EXCOV ALGORITHM

Maximizing the GML objective function (3.1) by an exhaustive search is prohibitively complex because we need to determine the ML estimate of the variance components $\widehat{\boldsymbol{\rho}}_A$ for each of 2^m candidates of the index set A. In this section, we describe our EXCOV method that approximately maximizes (3.1). The basic idea of EXCOV is to interleave

- expansion and compression steps that modify the current estimate of the index set A by one element per step, with goal to find a more efficient A, and
- expectation-maximization (EM) steps that increase the marginal likelihood of the variance components for a fixed A, thereby approximating $\widehat{\boldsymbol{\rho}}_A$.

Throughout the EXCOV algorithm, which contains multiple cycles, we keep track of $\boldsymbol{\theta}^\star = \big(A^\star, \boldsymbol{\rho}^\star_{A^\star}\big)$ and $\boldsymbol{s}^\star$, the best estimate of $\boldsymbol{\theta}$ [yielding the largest $\mathrm{GL}(\boldsymbol{\theta})$] and corresponding signal estimate obtained *in the latest cycle*. We also keep

track of $\boldsymbol{\theta}^{\star\star} = \left(A^{\star\star}, \boldsymbol{\rho}_{A^{\star\star}}^{\star\star}\right)$ and $\boldsymbol{s}^{\star\star}$, denoting the best estimate of $\boldsymbol{\theta}$ and corresponding signal estimate obtained in the entire history of the algorithm, *including all cycles*.

We now describe the ExCoV algorithm:

Step 0 (Algorithm initialization): Initialize the signal estimate $\boldsymbol{s}^{(0)}$ using the minimum ℓ_2-norm estimate

$$\boldsymbol{s}^{(0)} = H^T \left(H^T H\right)^{-1} \boldsymbol{y} \tag{4.1}$$

and construct $A^{(0)}$ using the indices of the $m_{A^{(0)}}$ largest elements of $\boldsymbol{s}^{(0)}$. The simplest choice of $m_{A^{(0)}}$ is

$$m_{A^{(0)}} = 1 \tag{4.2a}$$

which is particularly appealing in large-scale problems; another choice that we utilize is

$$m_{A^{(0)}} \;\; = \;\; \left\lfloor \frac{N}{2\,\ln(m/N)} \right\rfloor \tag{4.2b}$$

motivated by the asymptotic results in [36, Sec. 7.6.2]. Then, $B^{(0)} = \mathcal{A} \setminus A^{(0)}$ and $m_{B^{(0)}} = m - m_{A^{(0)}}$. Set the initial $\mathrm{GL}(\boldsymbol{\theta}^{\star\star}) = -\infty$.

Step 1 (Cycle initialization): Set the iteration counter $p = 0$ and choose the initial variance component estimates $\boldsymbol{\rho}_{A^{(0)}}^{(0)} = \left(\boldsymbol{\delta}_{A^{(0)}}^{(0)}, (\gamma^2)^{(0)}, (\sigma^2)^{(0)}\right)$ as

$$(\sigma^2)^{(0)} \;\; = \;\; (\boldsymbol{y} - H_{A^{(0)}}\, \boldsymbol{s}_{A^{(0)}}^{(0)})^T\, C^{-1}\, (\boldsymbol{y} - H_{A^{(0)}}\, \boldsymbol{s}_{A^{(0)}}^{(0)})\,/\,N \tag{4.3a}$$

$$(\delta_{A^{(0)},i}^2)^{(0)} \;\; = \;\; 10\,(\sigma^2)^{(0)}\,/\,[H_{A^{(0)}}^T\, C^{-1}\, H_{A^{(0)}}]_{i,i}, \quad i = 1, 2, \ldots, m_{A^{(0)}} \tag{4.3b}$$

$$(\gamma^2)^{(0)} \;\; = \;\; \min_{i=1,2,\ldots,m_{A^{(0)}}} (\delta_{A^{(0)},i}^2)^{(0)}. \tag{4.3c}$$

This selection yields a diffuse signal-coefficient pdf in (2.4a). Set the initial $\boldsymbol{\theta}^{\star} = \left(A^{(0)}, \boldsymbol{\rho}_{A^{(0)}}^{(0)}\right)$ and $\boldsymbol{s}^{\star} = \boldsymbol{s}^{(0)}$.

Step 2 (Expansion): Determine the signal index $k \in B^{(p)}$ that corresponds to the component of $\boldsymbol{s}_{B^{(p)}}^{(p)}$ with the largest magnitude

$$k = \arg \max_{\kappa \in B^{(p)}} \mathrm{abs}(s_{\kappa}^{(p)}) \tag{4.4a}$$

move the index k from $B^{(p)}$ to $A^{(p)}$, yielding:

$$A^{(p+1)} \;\; = \;\; \{A^{(p)}, k\},\; B^{(p+1)} = B^{(p)} \backslash \{k\},\; m_{A^{(p+1)}} = m_{A^{(p)}} + 1,\; m_{B^{(p+1)}} = m_{B^{(p)}} - 1 \tag{4.4b}$$

and construct the new 'expanded' vector of distinct variance components $\boldsymbol{\delta}_{A^{(p+1)}}^{(p)} = [(\boldsymbol{\delta}_{A^{(p)}}^{(p)})^T, (\gamma^2)^{(p)}]^T$ and model-parameter set $\boldsymbol{\rho}_{A^{(p+1)}}^{(p)} = \left(\boldsymbol{\delta}_{A^{(p+1)}}^{(p)}, (\gamma^2)^{(p)}, (\sigma^2)^{(p)}\right)$.

Step 3 (EM): Apply one EM step described in Section IV-A for $A = A^{(p+1)}$ and previous $\boldsymbol{\rho}_{A}^{(p)} = \boldsymbol{\rho}_{A^{(p+1)}}^{(p)}$, yielding the updated model parameter estimates $\boldsymbol{\rho}_{A^{(p+1)}}^{(p+1)} = \left(\boldsymbol{\delta}_{A^{(p+1)}}^{(p+1)}, (\gamma^2)^{(p+1)}, (\sigma^2)^{(p+1)}\right)$ and signal estimate $\boldsymbol{s}^{(p+1)}$. Define $\boldsymbol{\theta}^{(p+1)} = \left(A^{(p+1)}, \boldsymbol{\rho}_{A^{(p+1)}}^{(p+1)}\right)$.

Step 4 (Update $\boldsymbol{\theta}^\star$): Check the condition

$$\mathrm{GL}(\boldsymbol{\theta}^{(p+1)}) > \mathrm{GL}\big(\boldsymbol{\theta}^\star\big). \tag{4.5}$$

If it holds, set $\boldsymbol{\theta}^\star = \boldsymbol{\theta}^{(p+1)}$ and $\boldsymbol{s}^\star = \boldsymbol{s}^{(p+1)}$; otherwise, keep $\boldsymbol{\theta}^\star$ and $\boldsymbol{s}^\star$ intact.

Step 5 (Stop expansion?): Check the condition

$$\mathrm{GL}(\boldsymbol{\theta}^{(p+1)}) < \min\Big\{\mathrm{GL}(\boldsymbol{\theta}^{(p+1-L)}), \frac{1}{L}\sum_{l=0}^{L-1}\mathrm{GL}(\boldsymbol{\theta}^{(p-l)})\Big\} \tag{4.6}$$

where L denotes the length of a *moving-average window*. If (4.6) does not hold, increment p by one and go back to Step 2; otherwise, if (4.6) holds, increment p by one and go to Step 6.

Step 6 (Compression): Find the smallest element $(\delta^2_{A^{(p)},i_{\min}})^{(p)}$ of $\boldsymbol{\delta}^{(p)}_{A^{(p)}} = \big[(\delta^2_{A^{(p)},1})^{(p)}, (\delta^2_{A^{(p)},2})^{(p)}, \ldots, (\delta^2_{A^{(p)},m_{A^{(p)}}})^{(p)}\big]^T$, where

$$i_{\min} = \arg\min_{i=1,2,\ldots,m_{A^{(p)}}} (\delta^2_{A^{(p)},i})^{(p)} \tag{4.7a}$$

and determine the signal index $k \in A^{(p)}$ that corresponds to this element; move k from $A^{(p)}$ to $B^{(p)}$, yielding:

$$A^{(p+1)} \;=\; A^{(p)}\backslash\{k\},\; B^{(p+1)} = \{B^{(p)}, k\},\; m_{A^{(p+1)}} = m_{A^{(p)}} - 1,\; m_{B^{(p+1)}} = m_{B^{(p)}} + 1 \tag{4.7b}$$

and construct the new 'compressed' vector of distinct variance components $\boldsymbol{\delta}^{(p)}_{A^{(p+1)}} = [(\delta^2_{A^{(p)},1})^{(p)}, \ldots, (\delta^2_{A^{(p)},i_{\min}-1})^{(p)}, (\delta^2_{A^{(p)},i_{\min}+1})^{(p)}, \ldots, (\delta^2_{A^{(p)},m_{A^{(p)}}})^{(p)}]^T$ and model-parameter set $\boldsymbol{\rho}^{(p)}_{A^{(p+1)}} = \big(\boldsymbol{\delta}^{(p)}_{A^{(p+1)}}, (\gamma^2)^{(p)}, (\sigma^2)^{(p)}\big)$.

Step 7 (EM): Apply one EM step from Section IV-A for $A = A^{(p+1)}$ and previous $\boldsymbol{\rho}^{(p)}_{A} = \boldsymbol{\rho}^{(p)}_{A^{(p+1)}}$, yielding the updated model parameter estimates $\boldsymbol{\rho}^{(p+1)}_{A^{(p+1)}} = \big(\boldsymbol{\delta}^{(p+1)}_{A^{(p+1)}}, (\gamma^2)^{(p+1)}, (\sigma^2)^{(p+1)}\big)$ and the signal estimate $\boldsymbol{s}^{(p+1)}$.

Step 8 (Update $\boldsymbol{\theta}^\star$): Check the condition (4.5). If it holds, set $\boldsymbol{\theta}^\star = \boldsymbol{\theta}^{(p+1)}$ and $\boldsymbol{s}^\star = \boldsymbol{s}^{(p+1)}$; otherwise, keep $\boldsymbol{\theta}^\star$ and $\boldsymbol{s}^\star$ intact.

Step 9 (Stop compression and complete cycle?) Check the condition (4.6). If (4.6) does not hold, increment p by one and go back to Step 6; otherwise, if it holds, complete the current cycle and go to Step 10.

Step 10 (Update $\boldsymbol{\theta}^{\star\star}$): Check the condition $\mathrm{GL}\big(\boldsymbol{\theta}^\star\big) > \mathrm{GL}\big(\boldsymbol{\theta}^{\star\star}\big)$. If it holds, set $\boldsymbol{\theta}^{\star\star} = \boldsymbol{\theta}^\star$ and $\boldsymbol{s}^{\star\star} = \boldsymbol{s}^\star$; otherwise, keep $\boldsymbol{\theta}^{\star\star}$ and $\boldsymbol{s}^{\star\star}$ intact.

Step 11 (Stop cycling?): If $A^{\star\star}$ has changed between two *consecutive* cycles, set $m_{A^{(0)}} = m_{A^{\star\star}}$, construct $A^{(0)}$ as the indices of $m_{A^{(0)}}$ largest-magnitude elements of

$$\boldsymbol{s}^{(0)} = \boldsymbol{s}^{\star\star} + H^T\,(HH^T)^{-1}\,(\boldsymbol{y} - H\,\boldsymbol{s}^{\star\star}) \tag{4.8}$$

and go back to Step 1; otherwise, terminate the ExCoV algorithm with the final signal estimate $\boldsymbol{s}^{\star\star}$.

If $H\,H^T = I_N$, computing $A^{(0)}$ using (4.8) can be viewed as a single *hard-thresholding step* in [24, eq. (10)]. Note that the minimum ℓ_2-norm estimate $H^T\,(H^T\,H)^{-1}\,\boldsymbol{y}$ is a special case of (4.8), with $\boldsymbol{s}^{\star\star}$ set to the zero vector.

Therefore, we are using hard-thresholding steps to initialize individual cycles as well as the entire algorithm. compare Steps 0 and 11.

One EXCOV cycle consists of an expansion sequence followed by a compression sequence. The stopping condition (4.6) for expansion or compression sequences utilizes a moving-average criterion to monitor the improvement of the objective function. EXCOV is fairly *insensitive* to the choice of the moving average window size L. The algorithm terminates when the latest cycle fails to find a distinct variance component support set that improves $\mathrm{GL}(\boldsymbol{\theta})$. Finally, EXCOV algorithm outputs the parameter and signal estimates having the highest $\mathrm{GL}(\boldsymbol{\theta})$. Parts (c) and (d) of Fig. 1 illustrate the final output of the EXCOV algorithm for the simulation scenario in Section V-A, where spikes with circles correspond to the signal elements belonging to the best index set $A^{\star\star}$ obtained upon completion of the EXCOV iteration.

A. An EM Step for Estimating the Variance Components For Fixed A

Assume that the index set A is fixed and that a previous variance-component estimate $\boldsymbol{\rho}_A^{(p)} = \big(\boldsymbol{\delta}_A^{(p)}, (\gamma^2)^{(p)}, (\sigma^2)^{(p)}\big)$ is available. In Appendix A, we treat the signal-coefficient vector $\boldsymbol{s}$ as the *missing (unobserved) data* and derive an EM step that yields a new set of variance-component estimates $\boldsymbol{\rho}_A^{(p+1)}$ satisfying

$$\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta})\big|_{\boldsymbol{\rho}_A = \boldsymbol{\rho}_A^{(p+1)}} \geq \ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta})\big|_{\boldsymbol{\rho}_A = \boldsymbol{\rho}_A^{(p)}} \tag{4.9}$$

see e.g. [37] and [38] for a general exposition on the EM algorithm and its properties. Note that $\boldsymbol{s}$ and $\boldsymbol{y}$ together make up the *complete data*. The EM step consists of computing the expected complete log-likelihood (E step):

$$\mathrm{E}_{\boldsymbol{s}\,|\boldsymbol{y},\boldsymbol{\theta}}\big[\ln p(\boldsymbol{s},\boldsymbol{y} \,|\, \boldsymbol{\theta}) \,|\, \boldsymbol{y}, (A, \boldsymbol{\rho}_A^{(p)})\big] \tag{4.10a}$$

and selecting the new variance-component estimates that maximize (4.10a) with respect to $\boldsymbol{\rho}_A$ (M step):

$$\boldsymbol{\rho}_A^{(p+1)} = \arg\max_{\boldsymbol{\rho}_A} \; \mathrm{E}_{\boldsymbol{s}\,|\boldsymbol{y},\boldsymbol{\theta}}\big[\ln p(\boldsymbol{s},\boldsymbol{y} \,|\, \boldsymbol{\theta}) \,|\, \boldsymbol{y}, (A, \boldsymbol{\rho}_A^{(p)})\big]. \tag{4.10b}$$

In the E step, we first compute

$$\boldsymbol{s}_A^{(p+1)} \;=\; \mathrm{E}_{\boldsymbol{s}\,|\boldsymbol{y},\boldsymbol{\theta}}\big[\boldsymbol{s}_A \,|\, \boldsymbol{y}, (A, \boldsymbol{\rho}_A^{(p)})\big] = D_A^{(p)}\, H_A^T\, P^{(p+1)}\, \boldsymbol{y} \tag{4.11a}$$

$$\boldsymbol{s}_B^{(p+1)} \;=\; \mathrm{E}_{\boldsymbol{s}\,|\boldsymbol{y},\boldsymbol{\theta}}\big[\boldsymbol{s}_B \,|\, \boldsymbol{y}, (A, \boldsymbol{\rho}_A^{(p)})\big] = (\gamma^2)^{(p)}\, H_B^T\, P^{(p+1)}\, \boldsymbol{y} \tag{4.11b}$$

then construct the *empirical Bayesian signal estimate*[3]

$$\boldsymbol{s}^{(p+1)} = [s_1^{(p+1)}, s_2^{(p+1)}, \ldots, s_m^{(p+1)}]^T = \mathrm{E}_{\boldsymbol{s}\,|\boldsymbol{y},\boldsymbol{\theta}}\big[\boldsymbol{s} \,|\, \boldsymbol{y}, (A, \boldsymbol{\rho}_A^{(p)})\big] \tag{4.11c}$$

[3]Here, $\mathrm{E}_{\boldsymbol{s}\,|\,\boldsymbol{y},\boldsymbol{\theta}}[\boldsymbol{s} \,|\, \boldsymbol{y}, \boldsymbol{\theta}]$ denotes the mean of the pdf $p(\boldsymbol{s} \,|\, \boldsymbol{y}, \boldsymbol{\theta})$, which is the Bayesian minimum mean-square error (MMSE) estimate of $\boldsymbol{s}$ for *known* $\boldsymbol{\theta}$ [35, Sec. 11.4]; it is also the *linear MMSE estimate* of $\boldsymbol{s}$ [35, Th. 11.1]. Hence, $\boldsymbol{s}^{(p+1)}$ in (4.11c) is an *empirical Bayesian estimate* of $\boldsymbol{s}$, with the variance components replaced with their pth-iteration estimates.

by interleaving $\boldsymbol{s}_A^{(p+1)}$ and $\boldsymbol{s}_B^{(p+1)}$ according to the index sets A and B, and, finally, compute

$$\Omega^{(p+1)} = \mathrm{cov}_{\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}\big[\boldsymbol{s}_A \,|\, \boldsymbol{y}, (A, \boldsymbol{\rho}_A^{(p)})\big] = D_A^{(p)} - D_A^{(p)}\, H_A^T\, P^{(p+1)}\, H_A\, D_A^{(p)} \tag{4.11d}$$

$$\begin{aligned}\xi^{(p+1)} &= \mathrm{E}_{\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}\big[\boldsymbol{s}_B^T \boldsymbol{s}_B \,|\, \boldsymbol{y}, (A, \boldsymbol{\rho}_A^{(p)})\big] \\ &= \|\boldsymbol{s}_B^{(p+1)}\|_{\ell_2}^2 + (\gamma^2)^{(p)}\big\{m_{_B} - (\gamma^2)^{(p)}\mathrm{tr}[P^{(p+1)}(H_B\, H_B^T)]\big\}\end{aligned} \tag{4.11e}$$

$$\begin{aligned}\zeta^{(p+1)} &= \mathrm{E}_{\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}\big[(\boldsymbol{y} - H\,\boldsymbol{s})^T\, C^{-1}\, (\boldsymbol{y} - H\,\boldsymbol{s}) \,|\, \boldsymbol{y}, (A, \boldsymbol{\rho}_A^{(p)})\big] \\ &= (\boldsymbol{y} - H\boldsymbol{s}^{(p+1)})^T C^{-1} (\boldsymbol{y} - H\boldsymbol{s}^{(p+1)}) + (\sigma^2)^{(p)}\big\{N - (\sigma^2)^{(p)}\mathrm{tr}[P^{(p+1)}C]\big\}\end{aligned} \tag{4.11f}$$

where

$$D_A^{(p)} = \mathrm{diag}\{(\delta_{A,1}^2)^{(p)}, (\delta_{A,2}^2)^{(p)}, \ldots, (\delta_{A,m_A}^2)^{(p)}\} \tag{4.11g}$$

$$P^{(p+1)} = [H_A\, D_A^{(p)}\, H_A^T + (\gamma^2)^{(p)}\, H_B\, H_B^T + (\sigma^2)^{(p)}\, C]^{-1}. \tag{4.11h}$$

In the M step, we update the variance components $\boldsymbol{\rho}_A$ as follows:

$$(\delta_{A,i}^2)^{(p+1)} = (s_{A,i}^{(p+1)})^2 + [\Omega^{(p+1)}]_{i,i}, \quad i = 1, 2, \ldots, m_A \tag{4.12a}$$

$$(\gamma^2)^{(p+1)} = \frac{1}{m_{_B}}\, \xi^{(p+1)} \tag{4.12b}$$

$$(\sigma^2)^{(p+1)} = \frac{1}{N}\, \zeta^{(p+1)}. \tag{4.12c}$$

Note that the term $H_B\, H_B^T$ in (4.11e) and (4.11h) is efficiently computed via the identity:

$$H_B\, H_B^T = HH^T - H_A\, H_A^T. \tag{4.13}$$

For white noise $C = I_N$ and full model $A = \mathcal{A}$ where γ^2 is dropped, our EM step reduces to the EM step under the SBL model in [18].

B. An Approximate ExCoV Scheme

The above ExCoV method requires matrix-matrix multiplications, which is prohibitively expensive in large-scale applications in terms of both storage and computational complexity. We now develop a large-scale approximate ExCoV scheme that can be implemented using matrix-vector multiplications only.

Our approximations are built upon the following assumptions:

$$C = I_N \tag{4.14a}$$

$$HH^T = I_N \tag{4.14b}$$

$$\gamma^2 = 0 \tag{4.14c}$$

where (4.14a) and (4.14b) imply white noise and orthogonal sensing matrix, respectively. When (4.14c) holds, $\boldsymbol{s}_B$ is zero with probability one, corresponding to the the strictly sparse signal model. Our approximate ExCoV scheme is the ExCoV scheme *simplified by employing the assumptions (4.14)*, with the following three modifications.

1) An Approximate EM Step: Under the assumptions (4.14), (4.11b) is not needed, and (4.11a) becomes

$$\boldsymbol{s}_A^{(p+1)} = [H_A^T\, H_A + (\sigma^2)^{(p)}\, (D_A^{(p)})^{-1}]^{-1}\, H_A^T\, \boldsymbol{y} \tag{4.15a}$$

where we have used the matrix inversion identity (A.1b). Note that (4.15a) can be implemented using the conjugate-gradient approach [39, Sec. 7.4], thus avoiding matrix inversion and requiring only matrix-vector multiplications. We approximate updates of the variance components in (4.12c) and (4.12a) by the following lower bounds:[4]

$$(\sigma^2)^{(p+1)} \;\approx\; \|\boldsymbol{y} - H_A\, \boldsymbol{s}_A^{(p+1)}\|_{\ell_2}^2 \,/\, N \tag{4.15b}$$

$$(\delta_{A,i}^2)^{(p+1)} \;\approx\; \max\Big\{(s_{A,i}^{(p+1)})^2, \frac{(\sigma^2)^{(p+1)}}{10\, h_{A,i}}\Big\}, \quad i = 1, 2, \ldots, m_A \tag{4.15c}$$

where $(s_{A,i}^{(p+1)})^2$ is a simple one-sample variance estimate of $\delta_{A,i}^2$ and the regularization term $(\sigma^2)^{(p+1)}/(10\, h_{A,i})$ in (4.15c) ensures numerical stability of the solution to (4.15a). In particular, this term ensures that the (i,i)th element of $(\sigma^2)^{(p)}\, (D_A^{(p)})^{-1}$ is smaller than or equal to ten times the corresponding element of $H_A^T\, H_A$ (for all i), see (4.15a).

2) An Approximate $\mathrm{GL}(\boldsymbol{\theta})$*:* We obtain an approximate $\mathrm{GL}(\boldsymbol{\theta})$ that avoids determinant computations in (3.2):

$$\begin{aligned}\mathrm{GL}_{\mathrm{app}}(A, \boldsymbol{\delta}_A, \sigma^2) = \tfrac{1}{2}\Big\{ &- N\, \ln(2\,\pi) - \ln\big(\frac{N - m_A}{2}\big) - (N - m_A - 2)\ln(\sigma^2) \\ &- \boldsymbol{y}^T\, \{I_N - H_A[H_A^T\, H_A + \sigma^2\, D_A^{-1}(\boldsymbol{\delta}_A)]^{-1} H_A^T\}\boldsymbol{y}/\sigma^2 - \sum_{i=1}^{m_A} \ln\Big[\frac{h_{A,i}^2}{2\,(\sigma^2 + h_{A,i}\,\delta_{A,i}^2)}\Big]\Big\}\end{aligned} \tag{4.16}$$

in which we have approximated $H_A^T\, H_A$ by a diagonal matrix:

$$H_A^T\, H_A \approx \mathrm{diag}\{h_{A,1}, h_{A,2}, \ldots, h_{A,m_A}\} \tag{4.17}$$

where

$$h_{A,i} = H_A^T(:, i)\, H_A(:, i) \tag{4.18}$$

See Appendix C for the derivation of (4.16).

3) A Modified Step 2 (Expansion): Since $\gamma^2 = 0$ and, therefore, $\boldsymbol{s}_B = \boldsymbol{0}_{m_B \times 1}$, we need a minor modification of Step 2 (Expansion) as follows. Determine the element k of the single variance index set $B^{(p)}$ that corresponds to the element of $H_{B^{(p)}}^T\, (\boldsymbol{y} - H_{A^{(p)}}\, \boldsymbol{s}_{A^{(p)}}^{(p)})$ with the largest magnitude; move k from $B^{(p)}$ to $A^{(p)}$ as described in (4.4b), yielding $A^{(p+1)}$ and $B^{(p+1)}$; finally, construct the new 'expanded' vector of distinct variance components $\boldsymbol{\delta}_{A^{(p+1)}}^{(p)}$ as

$$\boldsymbol{\delta}_{A^{(p+1)}}^{(p)} = \Big[(\boldsymbol{\delta}_{A^{(p)}}^{(p)})^T, \frac{(\sigma^2)^{(p)}}{H^T(:, k)\, H(:, k)}\Big]^T \tag{4.19}$$

where our choice of the initial variance estimate for the added element is such that the $(m_{A^{(p+1)}}, m_{A^{(p+1)}})$th element of $(\sigma^2)^{(p)}\, (D_A^{(p)})^{-1}$ and the corresponding element of $H_A^T\, H_A$ are equal, see (4.15a).

[4]The right-hand side of (4.15b) is less than or equal to the corresponding right-hand side of (4.12c); similarly, $(s_{A,i}^{(p+1)})^2$ on the right-hand side of (4.15c) is less than or equal to the corresponding right-hand side of (4.12a).

We now summarize the approximate EXCOV scheme. Run the same EXCOV steps under the assumptions (4.14), with the EM step replaced by the approximate EM step in (4.15a)–(4.15c), $\mathrm{GL}(\boldsymbol{\theta})$ evaluated by $\mathrm{GL}_{\mathrm{app}}(A, \boldsymbol{\delta}_A, \sigma^2)$, and Step 2 (Expansion) modified as described above.

C. Complexity and Memory Requirements of ExCoV and SBL

We discuss the complexity and memory requirements of our EXCOV and approximate EXCOV schemes and compare them with the corresponding requirements for the SBL method.

In its most efficient form, one step of the SBL iteration requires inverting an $N \times N$ matrix and multiplying matrices of sizes $m \times N$ and $N \times m$ respectively, see [18, eq. 17] and [19, eq. 5]. The complexity for the inversion is $\mathcal{O}(N^3)$ and the matrix multiplication demands $\mathcal{O}(Nm^2)$ operations. Therefore, keeping (2.2) in mind, we conclude that the overall complexity of each SBL step is $\mathcal{O}(Nm^2)$. Furthermore, the storage requirement of SBL is $\mathcal{O}(m^2)$.

The computation complexity of EXCOV lies in the EM updates and the same number of $\mathrm{GL}(\boldsymbol{\theta})$ evaluations (3.2). Extensive simulation experiments show that the number of EM steps in EXCOV is typically similar to if not fewer than the number of SBL iterations. For one EM step in EXCOV, the matrix inversion of size $N \times N$ in (4.11h) and matrix-matrix multiplication of sizes both at $N \times N$ dominate the complexity, which require $\mathcal{O}(N^3)$ operations. In terms of computing (3.2), the dominating factor is $\ln|P(\boldsymbol{\theta})|$, involving $\mathcal{O}(N^3)$ operations. Therefore, the complexity of one EM step and $\mathrm{GL}(\boldsymbol{\theta})$ evaluation in EXCOV is $\mathcal{O}(N^3)$. The sensing matrix H is the largest matrix EXCOV needs to store, requiring $\mathcal{O}(Nm)$ memory storage. The huge reduction in both complexity and storage compared with SBL is simply because EXCOV estimates much fewer parameters than SBL; the differences in the number of parameters and convergence speed are particularly significant in large-scale problems.

The approximate EXCOV scheme removes the two complexity bottlenecks of the exact EXCOV: the EM update and $\mathrm{GL}(\boldsymbol{\theta})$ are replaced by the approximate EM step and $\mathrm{GL}_{\mathrm{app}}(A, \boldsymbol{\delta}_A, \sigma^2)$ in (4.16). If we implement (4.15a) in the approximate EM step using the conjugate-gradient approach, the algorithm involves purely matrix-vector operation of sizes at most $N \times m$ and $m \times 1$. The complexity of one EM step is reduced from $\mathcal{O}(N^3)$ to $\mathcal{O}(N\,m)$. In large-scale applications, the sensing matrix H is typically not explicitly stored but instead appears in the function-handle form [for example, random DFT sensing matrix can be implemented via the fast Fourier transform (FFT)]. In this case, the storage of the approximate EXCOV scheme is just $\mathcal{O}(m)$.

V. NUMERICAL EXAMPLES

We apply the proposed methods to reconstruct one- and two-dimensional signals from compressive samples and compare their performance with the competing approaches.

Prior to applying the EXCOV schemes, we scale the measurements $\boldsymbol{y}$ by a positive constant c so that $\boldsymbol{y}^T\, C^{-1}\, \boldsymbol{y}/N = 1$; after completion of the EXCOV iterations, we scale the obtained signal estimates by $1/c$, thus removing the scaling

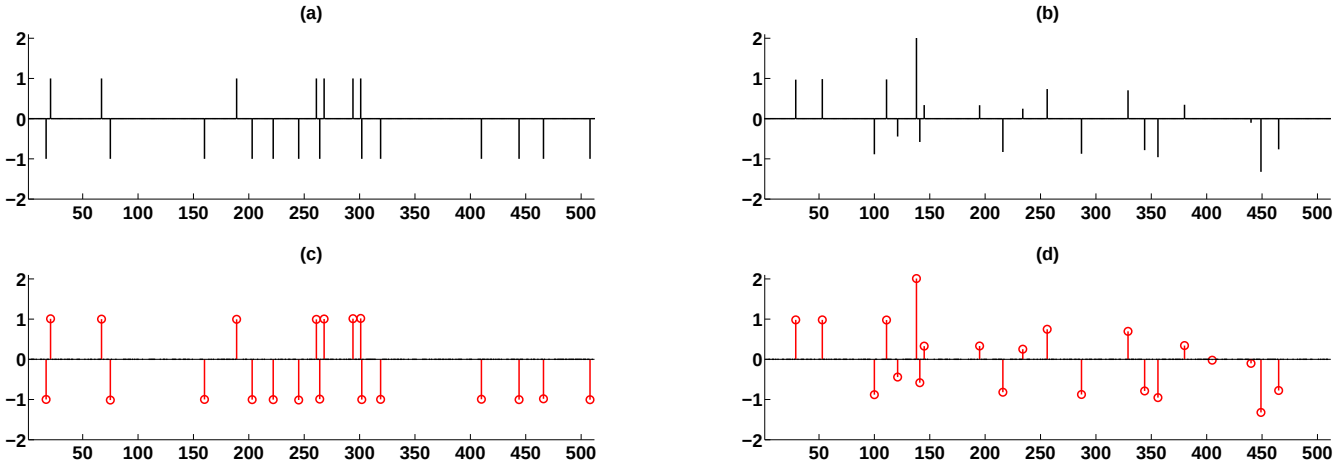

Fig. 1. Sparse signals with (a) binary and (b) Gaussian nonzero elements, respectively, and corresponding ExCoV reconstructions (c) and (d) from $N = 100$ noisy compressive samples, for noise variance 10^{-5}.

effect. This scaling, which we perform in all examples in this section, contributes to numerical stability and ensures that the estimates of σ^2 are less than or equal to one in all ExCoV iteration steps.

A. One-dimensional Signal Reconstruction

We generate the following standard test signals for sparse reconstruction methods, see also the simulation examples in [3], [5], [12], [20], and [26]. Consider sparse signals $\boldsymbol{s}$ of length $m = 512$, containing 20 *randomly located* nonzero elements. The nonzero components of $\boldsymbol{s}$ are independent, identically distributed (i.i.d.) random variables that are either

- *binary*, coming from the Rademacher distribution (i.e. taking values -1 or $+1$ with equal probability) or
- *Gaussian* with zero mean and unit variance

see parts (a) and (b) of Fig. 1 for sample signal realizations under the two models. In both cases, the variance of the nonzero elements of $\boldsymbol{s}$ is equal to one. The $N \times 1$ measurement vector $\boldsymbol{y}$ is generated using (2.1) with white noise having variance

$$\sigma^2 = 10^{-5}. \tag{5.1}$$

As in [12, Sec. IV.A] and the ℓ_1-magic suite of codes (available at http://www.l1-magic.org), the sensing matrices H are constructed by first creating an $N \times m$ matrix containing i.i.d. samples from the standard normal distribution and then orthonormalizing its rows, yielding $H\,H^T = I_N$.

Parts (c) and (d) of Fig. 1 present two examples of ExCoV reconstructions, for Gaussian and binary signals, respectively. Not surprisingly, the best index sets $A^{\star\star}$ obtained upon completion of the ExCoV iterations match well the true support sets of the signals, which is consistent with the essence of Theorem 1.

Our performance metric is the *average* mean-square error (MSE) of a signal estimate $\widehat{\boldsymbol{s}}$:

$$\mathrm{MSE}\{\widehat{\boldsymbol{s}}\} = \mathrm{E}_{\boldsymbol{y},\boldsymbol{s},H}[\|\widehat{\boldsymbol{s}} - \boldsymbol{s}\|_{\ell_2}^2]/m \tag{5.2}$$

computed using 2000 Monte Carlo trials, where *averaging* is performed over the random sensing matrices (H), the sparse signal $\boldsymbol{s}$ and the measurements $\boldsymbol{y}$. A simple benchmark of *poor performance* is the average MSE of the all-zero estimator, which is also the average signal energy: $\mathrm{MSE}\{\mathbf{0}_{m\times 1}\} = \mathrm{E}_{\boldsymbol{s},H}[\|\boldsymbol{s}\|_{\ell_2}^2]/m \approx 4\cdot 10^{-2}$.

We compare the following methods that represent state-of-the-art sparse reconstruction approaches of different types:

- the Bayesian compressive sensing (BCS) approach in [20], with a MATLAB implementation available at http://www.ece.duke.edu/~shji/BCS.html;
- the sparse Bayesian learning (SBL) method in [19, eq. (5)] which terminates when the squared norm of the difference of the signal estimates of two consecutive iterations is below $m\cdot 10^{-9}$;
- the second-order cone programming (SOCP) algorithm in [5] to solve the convex BPDN problem with the error-term size parameter ϵ chosen according to [5, eq. (3.1)] (as in the ℓ_1-magic package);
- the gradient-projection for sparse reconstruction (GPSR) method in [12, Sec. III.B] to solve the unconstrained version of the BPDN problem with the convergence threshold $\texttt{tolP} = 10^{-5}$ and regularization parameter $\tau = 0.01\,\|H^T\boldsymbol{y}\|_{\ell_\infty}$ (where tolP and τ have been manually tuned to achieve good reconstruction performance), see [12] and the GPSR suite of MATLAB codes at http://www.lx.it.pt/~mtf/GPSR;
- the normalized iterative hard thresholding (NIHT) method in [25] with the same convergence criterion as SBL, see the MATLAB implementation at http://www.see.ed.ac.uk/~tblumens/sparsify/sparsify.html;
- the standard and debiased compressive sampling matching pursuit algorithm in [16] (CoSaMP and CoSaMP-DB, respectively), with 300 iterations performed in each run.[5]
- our ExCoV and approximate ExCoV methods using $C = I_N$, averaging-window length $L = 10$, and initial value $m_{A^{(0)}}$ in (4.2b), with implementation available at http://home.eng.iastate.edu/~ald/ExCoV.htm;
- the clairvoyant least-squares (LS) signal estimator $\widehat{\boldsymbol{s}}_{\mathrm{LS}}$ for *known* locations of nonzero elements indexed by set A, obtained by setting $\widehat{\boldsymbol{s}}_{\mathrm{LS},A} = (H_A^T H_A)^{-1} H_A^T \boldsymbol{y}$ and the rest elements to zero (also discussed in [10, Sec. 1.2]), with average MSE

$$\mathrm{MSE}\{\widehat{\boldsymbol{s}}_{\mathrm{LS}}\} = \sigma^2\,\mathrm{E}_{A,H}\{\mathrm{tr}[(H_A^T H_A)^{-1}]\}/m. \tag{5.3}$$

(The above iterative methods were initialized using their default initial signal estimates, as specified in the references where they were introduced or implemented in the MATLAB functions provided by the authors.)

The CoSaMP and NIHT methods require knowledge of the number of nonzero elements in $\boldsymbol{s}$, and we use the true number 20 to implement both algorithms. SOCP needs the noise-variance parameter σ^2, and we use the true value 10^{-5} to implement it. In contrast, ExCoV is automatic and does not require prior knowledge about the signal or noise levels; furthermore, ExCoV does not employ a convergence tolerance level or threshold.

[5]Using more than 300 iterations does not improve the performance of the CoSaMP algorithm in our numerical examples. In the debiased CoSaMP, we compute the LS estimate of $\boldsymbol{s}$ using the sparse signal support obtained upon convergence of the CoSaMP algorithm.

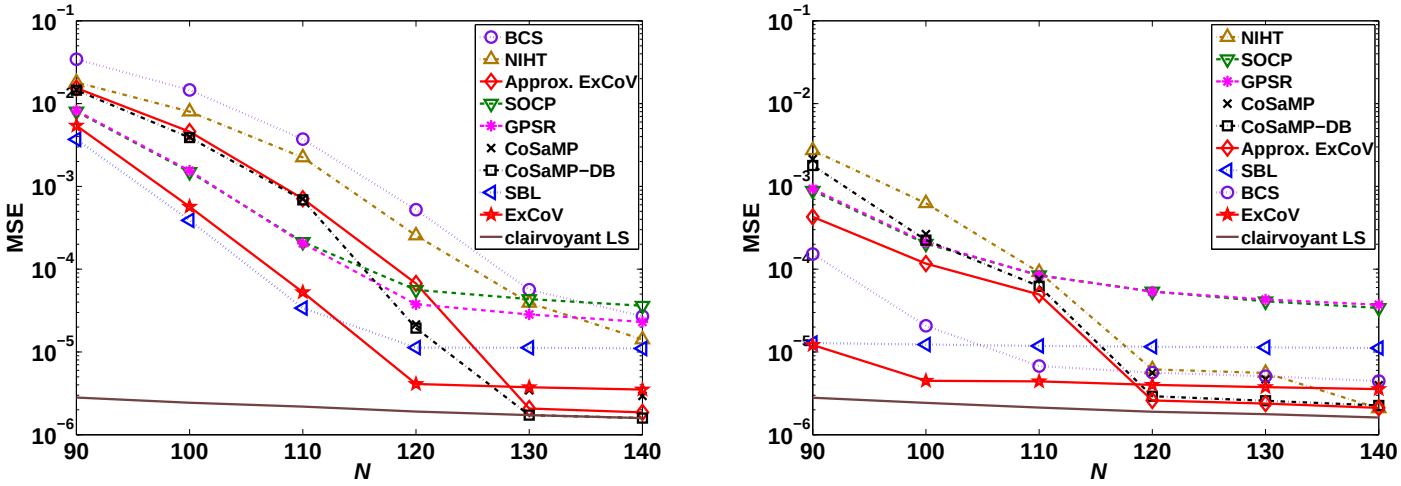

Fig. 2. Average MSEs of various estimators of $\boldsymbol{s}$ as functions of the number of measurements N, for (left) binary sparse signals and (right) Gaussian sparse signals, with noise variance equal to 10^{-5}.

Fig. 2 shows the average MSEs of the above methods as functions of the number of measurements N. For binary sparse signals and $90 \leq N \leq 110$, SBL achieves the smallest average MSE, closely followed by ExCoV; the convex methods SOCP and GPSR take the third place, with average MSE 1.5 to 3.9 times larger than that of ExCoV, see Fig. 2 (left). When N is sufficiently large ($N \geq 130$), ExCoV, approximate ExCoV, CoSaMP and CoSaMP-DB outperform SBL, with approximate ExCoV and CoSaMP-DB nearly attaining the average MSE of the clairvoyant LS method. Unlike CoSaMP and CoSaMP-DB, our ExCoV methods do not have the knowledge of the number of nonzero signal coefficients; yet, they approach the lower bound given by the clairvoyant LS estimator that knows the true signal support.

In this example, the numbers of iterations required by ExCoV and SBL methods are similar, but the CPU time of the former is much smaller than that of the latter. For example, when $N = 100$, ExCoV needs 155 EM steps on average and SBL converges in about 200 steps; however, the CPU time of SBL is 7.5 times that of ExCoV. Furthermore, the approximate ExCoV is much faster than both, consuming only about 3% of the CPU time of ExCoV for $N = 100$.

For Gaussian sparse signals and $N \leq 110$, ExCoV achieves the smallest average MSE, and SBL and BCS are the closest followers, see Fig. 2 (right). When N is sufficiently large ($N \geq 120$), approximate ExCoV, CoSaMP, CoSaMP-DB and NIHT catch up and achieve MSEs close to the clairvoyant LS lower bound.

For the same N, the average MSE (5.3) of clairvoyant LS is identical in the left- and right-hand sides of Fig. 2, since it is independent of the distribution of the non-zero signal elements. When N is small, the average MSEs for all methods and Gaussian sparse signals are much smaller than the binary counterparts, compare the left- and right-hand sides of Fig. 2. Indeed, it is well known that sparse binary signals are harder to estimate than other signals [26]. Interestingly, when there are enough measurements ($N \geq 130$), the average MSEs of most methods are similar for

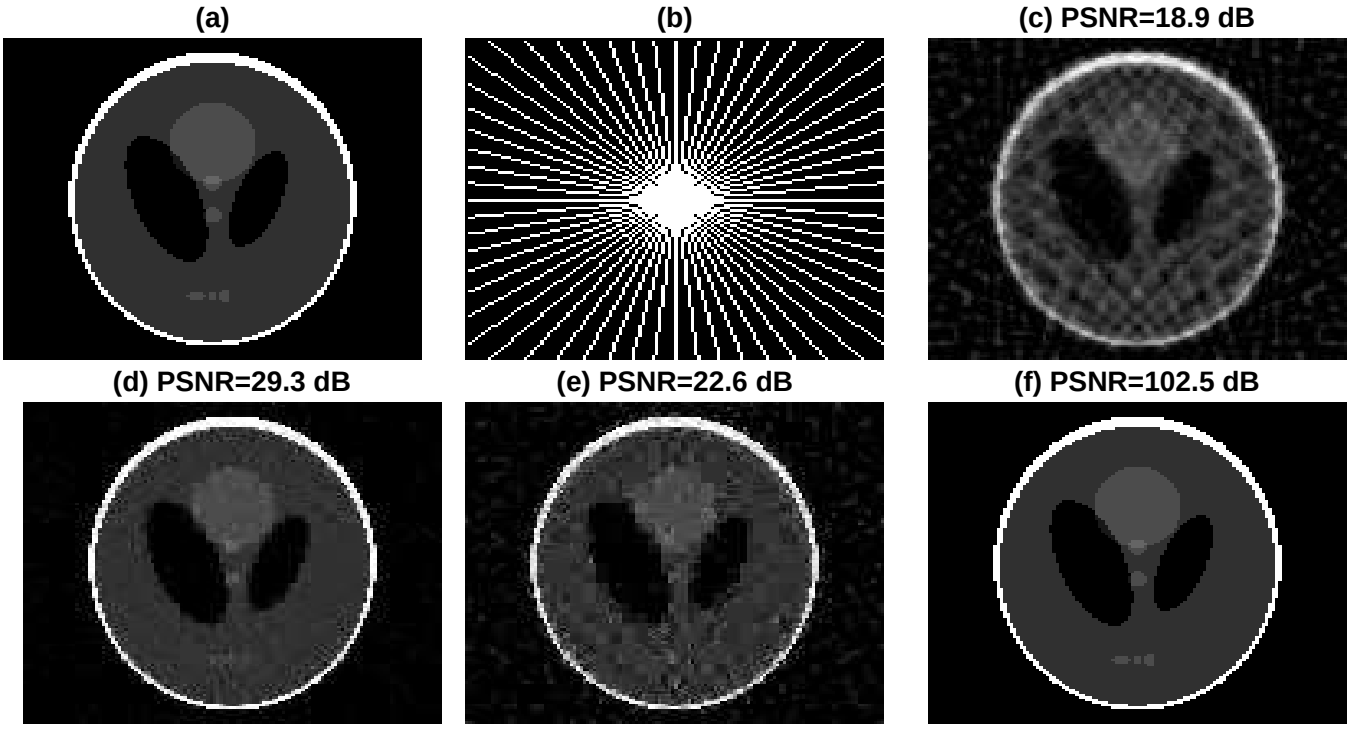

Fig. 3. (a) Size-128^2 Shepp-Logan phantom, (b) a star-shaped sampling domain in the frequency plane containing 30 radial lines, and reconstructions using (c) filtered back-projection, (d) NIHT, (e) GPSR-DB, and (f) approximate ExCoV schemes for the sampling pattern in (b).

binary and Gaussian sparse signals, with the exception of the BCS and NIHT schemes. Therefore, BCS and NIHT are sensitive to the distribution of the nonzero signal coefficients. Remarkably, for Gaussian sparse signals and sufficiently large N, NIHT almost attains the clairvoyant LS lower bound; yet, it does not perform well for binary sparse signals.

B. Two-dimensional Tomographic Image Reconstruction

Consider the reconstruction of the Shepp-Logan phantom of size $m = 128^2$ in Fig. 3 (a) from tomographic projections. The elements of $\boldsymbol{y}$ are 2-D discrete Fourier transform (DFT) coefficients of the image in Fig. 3 (a) sampled over a star-shaped domain, as illustrated in Fig. 3 (b); see also [5] and [25]. The sensing matrix is chosen as [2]

$$H = \Phi\, \Psi \tag{5.4}$$

with $N \times m$ *sampling matrix* Φ and $m \times m$ orthonormal sparsifying matrix Ψ constructed using selected rows of 2-D DFT matrix (yielding the corresponding 2-D DFT coefficients of the phantom image that are within the star-shaped domain) and inverse Haar wavelet transform matrix, respectively. Here, the rows of H are orthonormal, satisfying $H\, H^T = I_N$. The matrix H is not explicitly stored but instead implemented via FFT and wavelet function handle in MATLAB. The Haar wavelet coefficient vector $\boldsymbol{s}$ of the image in Fig. 3 (a) is sparse, with the number of nonzero elements equal to $1627 \approx 0.1\, m$. In the example in Fig. 3 (b), the samples are taken along 30 radial lines in the frequency plane, each containing 128 samples, which yields $N/m \approx 0.22$.

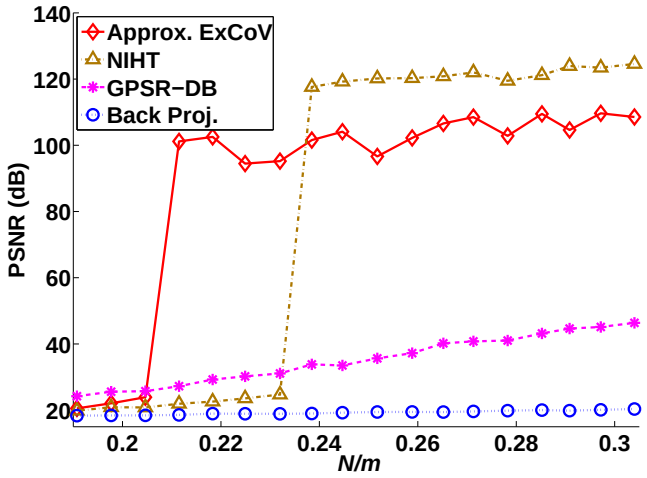

Fig. 4. PSNR as a function of the normalized number of measurements N/m, where the number of measurements changes by varying the number of radial lines in the star-shaped sampling domain.

Our performance metric is the peak signal-to-noise ratio (PSNR) of a wavelet coefficients estimate $\widehat{\boldsymbol{s}}$:

$$\text{PSNR (dB)} = 10 \log_{10} \left\{ \frac{[(\Psi \boldsymbol{s})_{\text{MAX}} - (\Psi \boldsymbol{s})_{\text{MIN}}]^2}{\|\widehat{\boldsymbol{s}} - \boldsymbol{s}\|_{\ell_2}^2 / m} \right\} \tag{5.5}$$

where $(\Psi \boldsymbol{s})_{\text{MIN}}$ and $(\Psi \boldsymbol{s})_{\text{MAX}}$ denote the smallest and largest elements of the image $\Psi \boldsymbol{s}$.

We compare the following representative reconstruction methods that are feasible for large-scale data:

- the standard filtered back-projection that corresponds to setting the unobserved DFT coefficients to zero and taking the inverse DFT, see [5];
- the debiased gradient-projection for sparse reconstruction method in [12, Sec. III.B] (labeled GPSR-DB) with convergence threshold $\texttt{tolP} = 10^{-5}$ and regularization parameter $\tau = 0.001 \, \|H^T \boldsymbol{y}\|_{\ell_\infty}$, both manually tuned to achieve good reconstruction performance;
- the NIHT method in [25], terminating when the squared norm of the difference of the signal estimates of two consecutive iterations is below $m \cdot 10^{-14}$;
- the approximate ExCoV method with averaging-window length $L = 100$ and initial value (4.2a), with signal estimation steps (4.15a) implemented using at most 300 conjugate-gradient steps.

Fig. 3 (c)–(f) present the reconstructed images from the 30 radial lines given in Fig. 3 (b) by the above methods. Approximate ExCoV manages to recover the original image almost perfectly, whereas the filtered back-projection method, NIHT and GPSR-DB have inferior reconstructions.

In Fig. 4, we vary the number of radial lines from 26 to 43, and, consequently, N/m from 0.19 to 0.31. We observe the sharp performance transition exhibited by approximate ExCoV at $N/m \approx 0.21$ (corresponding to 29 radial lines) very close to the theoretical minimum observation number, which is about twice the sparsity level $1627 \approx 0.1\, m$. Approximate ExCoV achieves almost perfect reconstruction with $N \approx 0.21\, m$ measurements. NIHT also exhibits

a sharp phase transition, but at $N/m \approx 0.24$ (corresponding to 33 radial lines), and GPSR-DB does not have a sharp phase transition in the range of N/m that we considered; rather, the PSNR of GPSR-DB improves with an approximately constant slope as we increase N/m.

VI. Concluding Remarks

We proposed a probabilistic model for sparse signal reconstruction and model selection. Our model generalizes the sparse Bayesian learning model, yielding a reduced parameter space. We then derived the GML function under the proposed probabilistic model that selects the most efficient signal representation making the best balancing between the accuracy of data fitting and compactness of the parameterization. We proved the equivalence of GML objective with the (P_0) optimization problem (1.1) and developed the ExCoV algorithm that searches for models with high GML objective function and provides corresponding empirical Bayesian signal estimates. ExCoV is automatic and does not require knowledge of the signal-sparsity or measurement-noise levels. We applied ExCoV to reconstruct one- and two-dimensional signals and compared it with the existing methods.

Further research will include analyzing the convergence of ExCoV, applying the GML rule to automate iterative hard thresholding algorithms (along the lines of [27]) and to select sparsifying matrices Ψ, and constructing GML-based distributed compressed sensing schemes for sensor networks, see also [41] and references therein for relevant work on compressed network sensing.

Appendix

We first present the EM step derivation (Appendix A) and then prove Theorem 1 (Appendix B), since some results from Appendix A are used in Appendix B; the derivation of $\mathrm{GL}_{\mathrm{app}}(A, \boldsymbol{\delta}_A, \sigma^2)$ in (4.16) is given in Appendix C.

Appendix A
EM Step Derivation

To derive the EM iteration (4.11)–(4.12), we repeatedly apply the matrix inversion lemma [40, eq. (2.22) at p. 424]:

$$(R + S\,T\,U)^{-1} = R^{-1} - R^{-1}\,S\,(T^{-1} + U R^{-1} S)^{-1}\,U\,R^{-1} \tag{A.1a}$$

and the following identity [40, p. 425]:

$$(R + S\,T\,U)^{-1}\,S\,T = R^{-1}\,S\,(T^{-1} + U\,R^{-1}\,S)^{-1} \tag{A.1b}$$

where R and T are invertible square matrices. The prior pdf (2.4a) can be written as

$$p(\boldsymbol{s}\,|\,\boldsymbol{\delta}_A, \gamma^2) = \mathcal{N}\big(\boldsymbol{s}\,;\,\mathbf{0}_{m\times 1}, D(\boldsymbol{\delta}_A, \gamma^2)\big) \tag{A.2}$$

where $D(\boldsymbol{\delta}_A, \gamma^2)$ is the $m \times m$ diagonal matrix with diagonal elements obtained by appropriately interleaving the variance components $\boldsymbol{\delta}_A$ and γ^2. Hence, $D_A(\boldsymbol{\delta}_A)$ and $D_B(\gamma^2)$ in (2.4b) are restrictions of the signal covariance matrix $D(\boldsymbol{\delta}_A, \gamma^2)$ to the index sets A and B. In particular, $D_A(\boldsymbol{\delta}_A)$ is the matrix of elements of $D(\boldsymbol{\delta}_A, \gamma^2)$ whose row and

column indices belong to the set A; similarly, $D_B(\gamma^2)$ is the matrix of elements of $D(\boldsymbol{\delta}_A, \gamma^2)$ whose row and column indices belong to B.

We treat the signal vector $\boldsymbol{s}$ as the *missing (unobserved) data*; then, the *complete-data log-likelihood function* of the measurements $\boldsymbol{y}$ and the missing data $\boldsymbol{s}$ given $\boldsymbol{\theta} = (A, \boldsymbol{\rho}_A)$ follows from (2.1) and (A.2):

$$\begin{aligned} \ln p(\boldsymbol{s}, \boldsymbol{y} \,|\, \boldsymbol{\theta}) \;\; &= \;\; \text{const} - \frac{N}{2}\, \ln(\sigma^2) - \frac{1}{2\,\sigma^2}\, (\boldsymbol{y} - H\,\boldsymbol{s})^T\, C^{-1}\, (\boldsymbol{y} - H\,\boldsymbol{s}) \\ &\quad -\tfrac{1}{2}\, \Big[\sum_{i=1}^{m_A} \ln(\delta_{A,i}^2) \Big] - \tfrac{1}{2}\, m_B\, \ln(\gamma^2) - \tfrac{1}{2}\, \boldsymbol{s}^T\, D^{-1}(\boldsymbol{\delta}_A, \gamma^2)\, \boldsymbol{s} \end{aligned} \tag{A.3}$$

where const denotes the terms not depending on $\boldsymbol{\theta}$ and $\boldsymbol{s}$. From (A.3), the conditional pdf of $\boldsymbol{s}$ given $\boldsymbol{y}$ and $\boldsymbol{\theta}$ is

$$p(\boldsymbol{s} \,|\, \boldsymbol{y}, \boldsymbol{\theta}) \propto \exp\big[-\tfrac{1}{2}\, (\boldsymbol{y} - H\,\boldsymbol{s})^T\, (\sigma^2\, C)^{-1}\, (\boldsymbol{y} - H\,\boldsymbol{s}) - \tfrac{1}{2}\, \boldsymbol{s}^T\, D^{-1}(\boldsymbol{\delta}_A, \gamma^2) \boldsymbol{s} \big] \tag{A.4}$$

yielding

$$p(\boldsymbol{s} \,|\, \boldsymbol{y}, \boldsymbol{\theta}) \;=\; \mathcal{N}\Big(\boldsymbol{s}\,;\, \big[D^{-1}(\boldsymbol{\delta}_A, \gamma^2) + H^T (\sigma^2\, C)^{-1} H\big]^{-1} H^T (\sigma^2\, C)^{-1} \boldsymbol{y},\, \big[D^{-1}(\boldsymbol{\delta}_A, \gamma^2) + H^T (\sigma^2\, C)^{-1} H\big]^{-1} \Big) \tag{A.5a}$$

$$\;=\; \mathcal{N}\Big(\boldsymbol{s}\,;\, D(\boldsymbol{\delta}_A, \gamma^2) H^T P(\boldsymbol{\theta}) \boldsymbol{y},\, D(\boldsymbol{\delta}_A, \gamma^2) - D(\boldsymbol{\delta}_A,\, \gamma^2)\, H^T\, P(\boldsymbol{\theta})\, H\, D(\boldsymbol{\delta}_A, \gamma^2) \Big) \tag{A.5b}$$

where $P(\boldsymbol{\theta}) = [H\, D(\boldsymbol{\delta}_A, \gamma^2)\, H^T + \sigma^2\, C]^{-1}$ was defined in (2.6b) and (A.5b) follows by applying (A.1a) and (A.1b). Then, (4.11a) and (4.11b) follow by setting $\boldsymbol{\theta} = (A, \boldsymbol{\rho}_A^{(p)})$ and restricting the mean vector in (A.5b) to the sub-vectors according to the index sets A and B. Similarly, (4.11d) follows by restricting the rows and columns of the covariance matrix in (A.5b) to a square sub-matrix according to index set A. Now,

$$\mathrm{E}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}(\boldsymbol{s}_B^T \boldsymbol{s}_B \,|\, \boldsymbol{y}, \boldsymbol{\theta}) \;\; = \;\; \mathrm{E}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}(\boldsymbol{s}_B \,|\, \boldsymbol{y}, \boldsymbol{\theta})^T\, \mathrm{E}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}(\boldsymbol{s}_B \,|\, \boldsymbol{y}, \boldsymbol{\theta}) + \mathrm{tr}\big[\mathrm{cov}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}(\boldsymbol{s}_B \,|\, \boldsymbol{y}, \boldsymbol{\theta})\big] \tag{A.6a}$$

$$\;\; = \;\; \|\mathrm{E}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}(\boldsymbol{s}_B \,|\, \boldsymbol{y}, \boldsymbol{\theta})\|_{\ell_2}^2 + \mathrm{tr}\big[\gamma^2 I_{m_B} - (\gamma^2)^2 H_B^T P(\boldsymbol{\theta}) H_B\big] \tag{A.6b}$$

where (A.6b) follows by restricting the rows and columns of the covariance matrix $\mathrm{cov}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}(\boldsymbol{s} \,|\, \boldsymbol{y}, \boldsymbol{\theta})$ in (A.5b) to the index set B. Setting $\boldsymbol{\rho}_A = \boldsymbol{\rho}_A^{(p)}$ leads to (4.11e). Similarly,

$$\begin{aligned} \mathrm{E}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}\big[(\boldsymbol{y} - H\,\boldsymbol{s})^T\, C^{-1}\, (\boldsymbol{y} - H\,\boldsymbol{s}) \,|\, \boldsymbol{y}, \boldsymbol{\theta}\big] \;\; &= \;\; \big[\boldsymbol{y} - H\, \mathrm{E}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}(\boldsymbol{s} \,|\, \boldsymbol{y}, \boldsymbol{\theta})\big]^T C^{-1} \big[\boldsymbol{y} - H\, \mathrm{E}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}(\boldsymbol{s} \,|\, \boldsymbol{y}, \boldsymbol{\theta})\big] \\ &\quad + \sigma^2\, \mathrm{tr}\big[H^T (\sigma^2\, C)^{-1} H\, \mathrm{cov}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}(\boldsymbol{s} \,|\, \boldsymbol{y}, \boldsymbol{\theta})\big] \end{aligned} \tag{A.7a}$$

where the second term simplifies by using (A.1b) and (A.5b): [see (2.6b)]:

$$\sigma^2\, \mathrm{tr}\big[H^T (\sigma^2\, C)^{-1} H\, \mathrm{cov}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}(\boldsymbol{s} \,|\, \boldsymbol{y}, \boldsymbol{\theta})\big] \;\; = \;\; \sigma^2\, \mathrm{tr}\big\{H^T (\sigma^2\, C)^{-1} H\, \big[D^{-1}(\boldsymbol{\delta}_A, \gamma^2) + H^T (\sigma^2\, C)^{-1} H\big]^{-1}\big\} \tag{A.7b}$$

$$\;\; = \;\; \sigma^2\, \mathrm{tr}\big\{[H\, D(\boldsymbol{\delta}_A, \gamma^2)\, H^T + \sigma^2\, C]^{-1} H D(\boldsymbol{\delta}_A, \gamma^2) H^T\big\} \tag{A.7c}$$

$$\;\; = \;\; \sigma^2\, \mathrm{tr}\big\{I_N - \sigma^2 P(\boldsymbol{\theta})\, C\big\} \tag{A.7d}$$

and (4.11f) follows by setting $\boldsymbol{\rho}_A = \boldsymbol{\rho}_A^{(p)}$. This concludes the derivation of the E step (4.11). The M step (4.12) easily follows by setting the derivatives of $\mathrm{E}_{\,\boldsymbol{s}\,|\,\boldsymbol{y},\,\boldsymbol{\theta}}\big[\ln p(\boldsymbol{s}, \boldsymbol{y} \,|\, \boldsymbol{\theta}) \,|\, \boldsymbol{y}, (A, \boldsymbol{\rho}_A^{(p)})\big]$ with respect to the variance components $\boldsymbol{\rho}_A = \big(\boldsymbol{\delta}_A, \gamma^2, \sigma^2\big)$ to zero.

APPENDIX B
PROOF OF THEOREM 1

We first prove a few useful lemmas.

Lemma 1: Consider an index set $A \subset \{1, 2, \ldots, m\}$ with cardinality $m_A \leq N$, defining distinct signal variance components. Assume that the URP condition (1) holds, distinct variance components are all positive, and the single variance for $B = \mathcal{A} \setminus A$ is zero, i.e. $\boldsymbol{\delta}_A \succ \mathbf{0}_{m_A}$ and $\gamma^2 = 0$, implying that A is the set of indices corresponding to all positive signal variance components. Then, the following hold:

$$\lim_{\sigma^2 \searrow 0} H_A^T \, P(\boldsymbol{\theta}) \;=\; D_A^{-1/2}(\boldsymbol{\delta}_A) \left[C^{-1/2} H_A \, D_A^{1/2}(\boldsymbol{\delta}_A)\right]^\dagger C^{-1/2} \tag{B.1a}$$

$$\lim_{\sigma^2 \searrow 0} H_A^T \, P(\boldsymbol{\theta}) \, H_A \;=\; D_A^{-1}(\boldsymbol{\delta}_A) \tag{B.1b}$$

$$\lim_{\sigma^2 \searrow 0} \sigma^2 \, P(\boldsymbol{\theta}) \;=\; C^{-1/2} \, \Pi^\perp(C^{-1/2} H_A) \, C^{-1/2} \tag{B.1c}$$

$$\lim_{\sigma^2 \searrow 0} \frac{\ln |P(\boldsymbol{\theta})|}{\ln(1/\sigma^2)} \;=\; N - m_A \tag{B.1d}$$

where $P(\boldsymbol{\theta})$ was defined in (2.6b) and, since $D_A(\boldsymbol{\delta}_A)$ is a diagonal matrix, $D_A^{1/2}(\boldsymbol{\delta}_A) = \mathrm{diag}\{\delta_{A,1}, \delta_{A,2}, \ldots, \delta_{A,m_A}\}$, $\delta_{A,i} = (\delta_{A,i}^2)^{1/2}$, $i = 1, 2, \ldots, m_A$.

Proof: Using (2.6b) and setting $\gamma^2 = 0$ leads to

$$\begin{aligned}\lim_{\sigma^2 \searrow 0} H_A^T \, P(\boldsymbol{\theta}) &= \lim_{\sigma^2 \searrow 0} H_A^T \left[H_A \, D_A(\boldsymbol{\delta}_A) \, H_A^T + \sigma^2 \, C\right]^{-1} \\ &= \lim_{\sigma^2 \searrow 0} D_A^{-1/2}(\boldsymbol{\delta}_A) \left[C^{-1/2} \, H_A \, D_A^{1/2}(\boldsymbol{\delta}_A)\right]^T \left[C^{-1/2} \, H_A \, D_A(\boldsymbol{\delta}_A) \, H_A^T \, C^{-1/2} + \sigma^2 \, I_N\right]^{-1} C^{-1/2}\end{aligned}$$

and (B.1a) follows by using the limiting form of the Moore-Penrose inverse [40, Th. 20.7.1]. Using (B.1a), we have

$$\lim_{\sigma^2 \searrow 0} H_A^T \, P(\boldsymbol{\theta}) \, H_A = D_A^{-1/2}(\boldsymbol{\delta}_A) \left[C^{-1/2} H_A \, D_A^{1/2}(\boldsymbol{\delta}_A)\right]^\dagger \left[C^{-1/2} H_A \, D_A^{1/2}(\boldsymbol{\delta}_A)\right] D_A^{-1/2}(\boldsymbol{\delta}_A)$$

and (B.1b) follows by noting that $m_A \leq N$ and $C^{-1/2} H_A \, D_A^{1/2}(\boldsymbol{\delta}_A)$ has full column rank m_A due to URP, see also [40, Th. 20.5.1]. Now, apply (A.1a):

$$\lim_{\sigma^2 \searrow 0} \sigma^2 P(\boldsymbol{\theta}) = \lim_{\sigma^2 \searrow 0} C^{-1/2} \big[I_N - C^{-1/2} H_A \left(\sigma^2 D_A^{-1}(\boldsymbol{\delta}_A) + H_A^T C^{-1} H_A\right)^{-1}\big] C^{-1/2}$$

and notice that $\left(H_A^T C^{-1} H_A\right)^{-1}$ exists due to $m_A \leq N$ and URP condition; (B.1c) then follows. Finally,

$$\begin{aligned}\ln |P(\boldsymbol{\theta})| &= -\ln |H_A \, D_A(\boldsymbol{\delta}_A) \, H_A^T + \sigma^2 \, C| = -\ln |\sigma^2 \, C| - \ln |H_A^T \, C^{-1} \, H_A \, D_A(\boldsymbol{\delta}_A)/\sigma^2 + I_{m_A}| \\ &= (N - m_A) \, \ln(1/\sigma^2) - \ln |C| - \ln |H_A^T \, C^{-1} \, H_A \, D_A(\boldsymbol{\delta}_A) + \sigma^2 \, I_{m_A}|\end{aligned}$$

where the last term is finite when $m_A \leq N$ and URP condition (1) holds, and (B.1d) follows. □

Under the conditions of Lemma 1 and if $m_A < N$, $P(\boldsymbol{\theta})$ is unbounded as $\sigma^2 \searrow 0$. Eqs. (B.1a)–(B.1c) show that multiplying $P(\boldsymbol{\theta})$ by H_A or σ^2 leads to bounded limiting expressions as $\sigma^2 \searrow 0$. When $m_A < N$, $\ln |P(\boldsymbol{\theta})|$ behaves as $(N - m_A) \, \ln(1/\sigma^2)$ as $\sigma^2 \searrow 0$, see (B.1d); the smaller m_A, the quicker $\ln |P(\boldsymbol{\theta})|$ grows to infinity.

We now examine $\boldsymbol{y}^T\,P(\boldsymbol{\theta})\,\boldsymbol{y}$ and the signal estimate [see (A.5b)]

$$\mathrm{E}_{\boldsymbol{s}\,|\,\boldsymbol{\theta},\boldsymbol{y}}[\boldsymbol{s}_A\,|\,\boldsymbol{y},\boldsymbol{\theta}] = D_A(\boldsymbol{\delta}_A)\,H_A^T\,P(\boldsymbol{\theta})\,\boldsymbol{y} \tag{B.2}$$

for the cases where the index set A *does* and *does not* includes the (P_0)-optimal support $A^\diamond$.

Lemma 2: As in Lemma 1, we assume that the URP condition (1) holds and $m_A \leq N$, $\boldsymbol{\delta}_A \succ \mathbf{0}_{m_A}$, and $\gamma^2 = 0$, implying that A is the set of indices corresponding to all positive signal variance components.

(a) If A includes the (P_0)-optimal support $A^\diamond$ ($A \supseteq A^\diamond$), then

$$\lim_{\sigma^2 \searrow 0} \boldsymbol{y}^T\,P(\boldsymbol{\theta})\,\boldsymbol{y} \;=\; (\boldsymbol{s}^\diamond_{A^\diamond})^T\,D^{-1}_{A^\diamond}(\boldsymbol{\delta}_{A^\diamond})\,\boldsymbol{s}^\diamond_{A^\diamond} \tag{B.3a}$$

$$\lim_{\sigma^2 \searrow 0} \mathrm{E}_{\boldsymbol{s}\,|\,\boldsymbol{y},\boldsymbol{\theta}}[\boldsymbol{s}_A\,|\,\boldsymbol{y},\boldsymbol{\theta}] \;=\; \boldsymbol{s}^\diamond_A. \tag{B.3b}$$

(b) If A does not include the (P_0)-optimal support $A^\diamond$ ($A \not\supseteq A^\diamond$) and $\mathrm{card}(A^\diamond \cup A) \leq N$, then

$$\lim_{\sigma^2 \searrow 0} \sigma^2\,\boldsymbol{y}^T\,P(\boldsymbol{\theta})\,\boldsymbol{y} = \|\Pi^\perp(C^{-1/2}\,H_A)\,C^{-1/2}\,H_{A^\diamond\cap B}\,\boldsymbol{s}^\diamond_{A^\diamond\cap B}\|^2_{\ell_2} > 0. \tag{B.3c}$$

Proof: $A \supseteq A^\diamond$ implies that the elements of $\boldsymbol{s}^\diamond$ with indices in $A\backslash A^\diamond$ are zero; consequently,

$$\boldsymbol{y} = H_{A^\diamond}\,\boldsymbol{s}^\diamond_{A^\diamond} = H_A\,\boldsymbol{s}^\diamond_A \tag{B.4}$$

and (B.3a)–(B.3b) follow by using (B.4), (B.1b), and (B.2).

We now show part (b) where $A \not\supseteq A^\diamond$. Observe that, when $\gamma^2 = 0$,

$$\begin{aligned}\boldsymbol{y}^T\,P(\boldsymbol{\theta})\,\boldsymbol{y} &= (\boldsymbol{s}^\diamond_{A^\diamond})^T\,H^T_{A^\diamond}\,P(\boldsymbol{\theta})\,H_{A^\diamond}\,\boldsymbol{s}^\diamond_{A^\diamond} = (\boldsymbol{s}^\diamond_{A^\diamond\cup A})^T\,H^T_{A^\diamond\cup A}\,P(\boldsymbol{\theta})\,H_{A^\diamond\cup A}\,\boldsymbol{s}^\diamond_{A^\diamond\cup A}\\ &= (\boldsymbol{s}^\diamond_A)^T\,H^T_A\,P(\boldsymbol{\theta})\,H_A\,\boldsymbol{s}^\diamond_A + 2\,(\boldsymbol{s}^\diamond_A)^T\,H^T_A\,P(\boldsymbol{\theta})\,H_{A^\diamond\cap B}\,\boldsymbol{s}^\diamond_{A^\diamond\cap B} + (\boldsymbol{s}^\diamond_{A^\diamond\cap B})^T\,H^T_{A^\diamond\cap B}\,P(\boldsymbol{\theta})\,H_{A^\diamond\cap B}\,\boldsymbol{s}^\diamond_{A^\diamond\cap B}\end{aligned} \tag{B.5}$$

which follows by using (B.4) and partitioning $A^\diamond \cup A$ into A and $A^\diamond \cap B$. The first two terms in (B.5) are finite at $\sigma^2 = 0$, which easily follows by employing (B.1a) and (B.1b). Then, the equality in (B.3c) follows by using (B.1c). We now show that (B.3c) is positive by contradiction. The URP property of H and the assumption that $\mathrm{card}(A^\diamond \cup A) \leq N$ imply that the columns of $H_{A^\diamond\cup A}$ are linearly independent. Since $\boldsymbol{s}_{A^\diamond\cap B}$ is a nonzero vector and columns of $H_{A^\diamond\cap B}$ are linearly independent, $C^{-1/2}H_{A^\diamond\cap B}\,\boldsymbol{s}^\diamond_{A^\diamond\cap B}$ is a nonzero vector. If (B.3c) is zero, then $C^{-1/2}H_{A^\diamond\cap B}\,\boldsymbol{s}^\diamond_{A^\diamond\cap B}$ belongs to the column space of $C^{-1/2}H_A$, which contradicts the fact that the columns of $C^{-1/2}H_{A^\diamond\cup A}$ are linearly independent.

□

Lemma 2 examines the behavior of $\boldsymbol{y}^T\,P(\boldsymbol{\theta})\,\boldsymbol{y}$ and the signal estimate (B.2) as noise variance shrinks to zero. Clearly, $A \supseteq A^\diamond$ is desirable and, in contrast, there is a severe penalty if $A \not\supseteq A^\diamond$. Under the assumptions of Lemma 2, $\boldsymbol{y}^T\,P(\boldsymbol{\theta})\,\boldsymbol{y}$ [which is an important term in the log-likelihood function (2.6c)] is finite when A includes all elements of $A^\diamond$, see (B.3a); in contrast, when A misses any index from $A^\diamond$, $\boldsymbol{y}^T\,P(\boldsymbol{\theta})\,\boldsymbol{y}$ grows hyperbolically with σ^2 as $\sigma^2 \searrow 0$, see (B.3c). Furthermore, if A includes $A^\diamond$, (B.3b) holds regardless of the specific values of $\boldsymbol{\delta}_A$ *provided that* they

are positive; hence, the signal estimate $\mathrm{E}_{\boldsymbol{s}\,|\,\boldsymbol{y},\boldsymbol{\theta}}[\boldsymbol{s}_A\,|\,\boldsymbol{y},\boldsymbol{\theta}]$ will be (P_0)-optimal even if the variance components are inaccurate. The next lemma studies the behavior of the Fisher information term of the GML function.

Lemma 3: For any distinct-variance index set $A \subseteq \{1,2,\ldots,m\}$, define the index set of positive variance components in A:

$$A^+(\boldsymbol{\delta}_A) \;\stackrel{\triangle}{=}\; \{i \in A \,:\, [D(\boldsymbol{\delta}_A,\gamma^2)]_{i,i} > 0\} \tag{B.6a}$$

with cardinality

$$m_{A^+} \;\stackrel{\triangle}{=}\; \mathrm{card}(A^+(\boldsymbol{\delta}_A)) \leq m_A. \tag{B.6b}$$

Assume that the URP and Fisher-information conditions (1) and (2) hold.

(a) If $\gamma^2 = 0$, then

$$\lim_{\sigma^2 \searrow 0} \frac{\ln|\mathcal{I}(\boldsymbol{\theta})|}{\ln(1/\sigma^2)} = \left\{ \begin{array}{cc} 2\,(m_A - m_{A^+} + 1), & \text{if } m_{A^+} < N \\ 0, & \text{if } m_{A^+} \geq N \end{array} \right. . \tag{B.7a}$$

(b) If $\gamma^2 > 0$, then

$$\lim_{\sigma^2 \searrow 0} \frac{\ln|\mathcal{I}(\boldsymbol{\theta})|}{\ln(1/\sigma^2)} = \left\{ \begin{array}{cc} 2\,(m_A - m_{A^+}), & \text{if } m_{A^+} + m_B < N \\ 0, & \text{if } m_{A^+} + m_B \geq N \end{array} \right. . \tag{B.7b}$$

Proof: Without loss of generality, let $A^+(\boldsymbol{\delta}_A) = A^+ = \{1,2,\ldots,m_{A^+}\}$ and block partition $\mathcal{I}_{\boldsymbol{\delta}_A,\boldsymbol{\delta}_A}(\boldsymbol{\theta})$ as:

$$\mathcal{I}_{\boldsymbol{\delta}_A,\boldsymbol{\delta}_A}(\boldsymbol{\theta}) = \left[\begin{array}{cc} \mathcal{I}_{\boldsymbol{\delta}_{A^+},\boldsymbol{\delta}_{A^+}}(\boldsymbol{\theta}) & \mathcal{I}_{\boldsymbol{\delta}_{A^+},\boldsymbol{\delta}_{A\setminus A^+}}(\boldsymbol{\theta}) \\ \mathcal{I}^T_{\boldsymbol{\delta}_{A^+},\boldsymbol{\delta}_{A\setminus A^+}}(\boldsymbol{\theta}) & \mathcal{I}_{\boldsymbol{\delta}_{A\setminus A^+},\boldsymbol{\delta}_{A\setminus A^+}}(\boldsymbol{\theta}) \end{array} \right]. \tag{B.8}$$

We first show part (a), where $\gamma^2 = 0$ and, therefore, $P(\boldsymbol{\theta}) = [H_{A^+}\,D_{A^+}(\boldsymbol{\delta}_{A^+})\,H^T_{A^+} + \sigma^2\,I_N]^{-1}$. When $m_{A^+} \geq N$, the URP property of H implies that $P(\boldsymbol{\theta})$ and $\mathcal{I}(\boldsymbol{\theta})$ are finite matrices and

$$\lim_{\sigma^2 \searrow 0} \frac{\ln|\mathcal{I}(\boldsymbol{\theta})|}{\ln(1/\sigma^2)} = 0. \tag{B.9}$$

Consider now the case where $m_{A^+} < N$ and, consequently, $P(\boldsymbol{\theta})$ is unbounded as $\sigma^2 \searrow 0$. Applying Lemma 1 to the index set A^+ implies that multiplying $P(\boldsymbol{\theta})$ by H_{A^+} or σ^2 leads to bounded expressions; in particular, we obtain

$$\lim_{\sigma^2 \searrow 0} \mathcal{I}_{\boldsymbol{\delta}_{A^+},\boldsymbol{\delta}_{A^+}}(\boldsymbol{\theta}) \;=\; \lim_{\sigma^2 \searrow 0} \tfrac{1}{2}\,[H^T_{A^+} P(\boldsymbol{\theta}) H_{A^+}]^{\odot 2} = \tfrac{1}{2} D^{-2}_{A^+}(\boldsymbol{\delta}_{A^+}) \tag{B.10a}$$

$$\lim_{\sigma^2 \searrow 0} \mathcal{I}_{\boldsymbol{\delta}_{A^+},\boldsymbol{\delta}_{A\setminus A^+}}(\boldsymbol{\theta}) \;=\; \tfrac{1}{2}\left\{D^{-1/2}_{A^+}(\boldsymbol{\delta}_{A^+})\left[C^{-1/2} H_{A^+}\,D^{1/2}_{A^+}(\boldsymbol{\delta}_{A^+})\right]^\dagger C^{-1/2} H_{A\setminus A^+}\right\}^{\odot 2} \tag{B.10b}$$

$$\lim_{\sigma^2 \searrow 0} (\sigma^2)^2\,\mathcal{I}_{\boldsymbol{\delta}_{A\setminus A^+},\boldsymbol{\delta}_{A\setminus A^+}}(\boldsymbol{\theta}) \;=\; \tfrac{1}{2}\,[H^T_{A\setminus A^+}\,C^{-1/2}\,\Pi^\perp(C^{-1/2}H_{A^+})\,C^{-1/2}\,H_{A\setminus A^+}]^{\odot 2}$$

$$\lim_{\sigma^2 \searrow 0} \mathcal{I}_{\boldsymbol{\delta}_{A^+},\gamma^2}(\boldsymbol{\theta}) \;=\; \tfrac{1}{2}\;\mathrm{diag}\left\{D^{-1/2}_{A^+}(\boldsymbol{\delta}_{A^+})\,[C^{-1/2}H_{A^+}D^{1/2}_{A^+}(\boldsymbol{\delta}_{A^+})]^\dagger\,C^{-1/2}\,H_B\right.$$
$$\left.\cdot\left[D^{-1/2}_{A^+}(\boldsymbol{\delta}_{A^+})\,[C^{-1/2}\,H_{A^+}\,D^{1/2}_{A^+}(\boldsymbol{\delta}_{A^+})]^\dagger\,C^{-1/2}\,H_B\right]^T\right\} \tag{B.10c}$$

$$\lim_{\sigma^2 \searrow 0} (\sigma^2)^2\,\mathcal{I}_{\boldsymbol{\delta}_{A\setminus A^+},\gamma^2}(\boldsymbol{\theta}) \;=\; \tfrac{1}{2}\;\mathrm{diag}\left\{H^T_{A\setminus A^+}\,C^{-1/2}\,\Pi^\perp(C^{-1/2}H_{A^+})\,C^{-1/2}\,H_B\right.$$
$$\left.\cdot\left[H^T_{A\setminus A^+}C^{-1/2}\Pi^\perp(C^{-1/2}H_{A^+})\,C^{-1/2}\,H_B\right]^T\right\} \tag{B.10d}$$

$$\lim_{\sigma^2 \searrow 0} (\sigma^2)^2\,\mathcal{I}_{\gamma^2,\gamma^2}(\boldsymbol{\theta}) \;=\; \tfrac{1}{2}\,\mathrm{tr}\left\{[C^{-1/2}\Pi^\perp(C^{-1/2}H_{A^+})\,C^{-1/2}\,H_B\,H^T_B]^2\right\} \tag{B.10e}$$

where the limits in (B.10a)-(B.10e) are all finite, see also (3.4).

We analyze the Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})$ and multiply by σ^2 all terms that contain $P(\boldsymbol{\theta})$ and are not guarded by H_{A^+}. In particular, multiplying the last $m_A - m_{A^+} + 1$ rows and columns of $\mathcal{I}(\boldsymbol{\theta})$ by σ^2 respectively leads to

$$\ln|\mathcal{I}(\boldsymbol{\theta})| = 2\,(m_A - m_{A^+} + 1)\ln(1/\sigma^2) + \ln\left|\begin{array}{ccc} \mathcal{I}_{\boldsymbol{\delta}_{A^+},\boldsymbol{\delta}_{A^+}}(\boldsymbol{\theta}) & \sigma^2\,\mathcal{I}_{\boldsymbol{\delta}_{A^+},\boldsymbol{\delta}_{A\setminus A^+}}(\boldsymbol{\theta}) & \sigma^2\,\mathcal{I}_{\boldsymbol{\delta}_{A^+},\gamma^2}(\boldsymbol{\theta}) \\ \sigma^2\,\mathcal{I}^T_{\boldsymbol{\delta}_{A^+},\boldsymbol{\delta}_{A\setminus A^+}}(\boldsymbol{\theta}) & (\sigma^2)^2\,\mathcal{I}_{\boldsymbol{\delta}_{A\setminus A^+},\boldsymbol{\delta}_{A\setminus A^+}}(\boldsymbol{\theta}) & (\sigma^2)^2\,\mathcal{I}_{\boldsymbol{\delta}_{A\setminus A^+},\gamma^2}(\boldsymbol{\theta}) \\ \sigma^2\,\mathcal{I}^T_{\boldsymbol{\delta}_{A^+},\gamma^2}(\boldsymbol{\theta}) & (\sigma^2)^2\,\mathcal{I}^T_{\boldsymbol{\delta}_{A\setminus A^+},\gamma^2}(\boldsymbol{\theta}) & (\sigma^2)^2\,\mathcal{I}_{\gamma^2,\gamma^2}(\boldsymbol{\theta}) \end{array}\right| \tag{B.11}$$

and (B.7a) follows.

We now show part (b), where $\gamma^2 > 0$ and $P(\boldsymbol{\theta}) = [H_{A^+}\,D_{A^+}(\boldsymbol{\delta}_{A^+})\,H^T_{A^+} + \gamma^2\,H_B\,H_B^2 + \sigma^2\,I_N]^{-1}$. When $m_{A^+} + m_B \geq N$, the URP property of H results in finite $P(\boldsymbol{\theta})$ and, therefore, $\mathcal{I}(\boldsymbol{\theta})$ is also finite, leading to

$$\lim_{\sigma^2\searrow 0}\frac{\ln|\mathcal{I}(\boldsymbol{\theta})|}{\ln(1/\sigma^2)} = 0. \tag{B.12}$$

When $m_{A^+} + m_B < N$, we have

$$\ln|\mathcal{I}(\boldsymbol{\theta})| = 2\,(m_A - m_{A^+})\ln(1/\sigma^2) + \ln\left|\begin{array}{ccc} \mathcal{I}_{\boldsymbol{\delta}_{A^+},\boldsymbol{\delta}_{A^+}}(\boldsymbol{\theta}) & \sigma^2\,\mathcal{I}_{\boldsymbol{\delta}_{A^+},\boldsymbol{\delta}_{A\setminus A^+}}(\boldsymbol{\theta}) & \mathcal{I}_{\boldsymbol{\delta}_{A^+},\gamma^2}(\boldsymbol{\theta}) \\ \sigma^2\,\mathcal{I}^T_{\boldsymbol{\delta}_{A^+},\boldsymbol{\delta}_{A\setminus A^+}}(\boldsymbol{\theta}) & (\sigma^2)^2\,\mathcal{I}_{\boldsymbol{\delta}_{A\setminus A^+},\boldsymbol{\delta}_{A\setminus A^+}}(\boldsymbol{\theta}) & \sigma^2\,\mathcal{I}_{\boldsymbol{\delta}_{A\setminus A^+},\gamma^2}(\boldsymbol{\theta}) \\ \mathcal{I}^T_{\boldsymbol{\delta}_{A^+},\gamma^2}(\boldsymbol{\theta}) & \sigma^2\,\mathcal{I}^T_{\boldsymbol{\delta}_{A\setminus A^+},\gamma^2}(\boldsymbol{\theta}) & \mathcal{I}_{\gamma^2,\gamma^2}(\boldsymbol{\theta}) \end{array}\right| \tag{B.13}$$

and (B.7b) follows by applying Lemma 1 for $A^+ \cup B$ and arguments analogous to those in part (a). □

From Lemma 3, we see that the Fisher information term of GML *penalizes* inclusion of zero variance components into index set A. In the following lemma, we analyze ML variance-component estimation for the full model $A = \mathcal{A}$.

Lemma 4: Consider the full model with $A = \mathcal{A}$ and empty B [see (2.3)], implying $\boldsymbol{\theta} = (\mathcal{A}, \boldsymbol{\rho}_{\mathcal{A}})$ and the variance-component parameter vector equal to $\boldsymbol{\rho}_{\mathcal{A}} = (\boldsymbol{\delta}, \sigma^2)$, where $\boldsymbol{\delta} = [\delta^2_{\mathcal{A},1}, \delta^2_{\mathcal{A},2}, \ldots, \delta^2_{\mathcal{A},m}]^T$. In this case, the log-likelihood function of the variance components is (2.6c) with $P(\boldsymbol{\theta}) = (H\ \mathrm{diag}\{\boldsymbol{\delta}\}\,H^T + \sigma^2\,C)^{-1}$. Assume that the URP and measurement number conditions (1) and (3) hold and consider all $\widehat{\boldsymbol{\rho}}_{\mathcal{A}} = (\widehat{\boldsymbol{\delta}}, \widehat{\sigma}^2)$ that satisfy

$$\mathcal{A}^+(\widehat{\boldsymbol{\delta}}) \;=\; \big\{i \in \mathcal{A} : \widehat{\delta}^2_{\mathcal{A},i} > 0\big\} = A^\diamond \tag{B.14a}$$

$$\widehat{\sigma}^2 \;=\; 0 \tag{B.14b}$$

where (B.14a) states that the support of $\widehat{\boldsymbol{\delta}} = [\widehat{\delta}^2_{\mathcal{A},1}, \widehat{\delta}^2_{\mathcal{A},2}, \ldots, \widehat{\delta}^2_{\mathcal{A},m}]^T$ is identical to the (P_0)-optimal support $A^\diamond$. Then, the log-likelihood $\ln p(\boldsymbol{y}\,|\,\boldsymbol{\theta})$ at $\boldsymbol{\delta} = \widehat{\boldsymbol{\delta}}$ grows proportionally to $\ln(1/\sigma^2)$ as σ^2 approaches $\widehat{\sigma}^2 = 0$, with speed

$$\lim_{\sigma^2\searrow 0}\frac{\ln p(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\ln(1/\sigma^2)}\Big|_{\boldsymbol{\delta}=\widehat{\boldsymbol{\delta}}} = \tfrac{1}{2}\,(N - m_{A^\diamond}). \tag{B.14c}$$

If $\sigma^2 > 0$, $p(\boldsymbol{y}\,|\,\boldsymbol{\theta})$ is always finite; therefore, it can become infinitely large only if $\sigma^2 = \widehat{\sigma}^2 = 0$. Among all choices of $\boldsymbol{\rho}_{\mathcal{A}}$ for which $p(\boldsymbol{y}\,|\,\boldsymbol{\theta})$ is infinitely large, those $\boldsymbol{\rho}_{\mathcal{A}} = \widehat{\boldsymbol{\rho}}_{\mathcal{A}}$ defined by (B.14a) and (B.14b) 'maximize' the likelihood in the sense that $\ln p(\boldsymbol{y}\,|\,\boldsymbol{\theta})$ grows to infinity at the fastest rate as $\sigma^2 \searrow 0$, quantified by (B.14c). Any choice of $\boldsymbol{\delta}$ different from $\widehat{\boldsymbol{\delta}}$ in (B.14a) *cannot* achieve this rate and, therefore, has a 'smaller' likelihood than $\widehat{\boldsymbol{\delta}}$ at $\sigma^2 = 0$.

Proof: Consider $\boldsymbol{\delta} = \widehat{\boldsymbol{\delta}}$ satisfying (B.14a), i.e. $\mathcal{A}^+(\widehat{\boldsymbol{\delta}}) = A^\diamond$. Applying (B.1d) in Lemma 1 and (B.3a) in Lemma 2 (a) for the index set $\mathcal{A}^+(\widehat{\boldsymbol{\delta}}) = A^\diamond$ yields (B.14c):

$$\lim_{\sigma^2 \searrow 0} \frac{\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta})}{\ln(1/\sigma^2)}\Big|_{\boldsymbol{\delta}=\widehat{\boldsymbol{\delta}}} = \lim_{\sigma^2 \searrow 0} \frac{-\frac{1}{2}\, N\, \ln(2\,\pi) + \frac{1}{2} \ln|P(\boldsymbol{\theta})| - \frac{1}{2} \boldsymbol{y}^T\, P(\boldsymbol{\theta})\, \boldsymbol{y}}{\ln(1/\sigma^2)}\Big|_{\boldsymbol{\delta}=\widehat{\boldsymbol{\delta}}} = \frac{N - m_{A^\diamond}}{2}. \tag{B.15}$$

We now examine the model parameters $\boldsymbol{\rho}_{\mathcal{A}}$ different from $\widehat{\boldsymbol{\rho}}_{\mathcal{A}}$ in (B.14). If $\sigma^2 > 0$, $P(\boldsymbol{\theta})$ is bounded and, therefore, the likelihood $p(\boldsymbol{y} \,|\, \boldsymbol{\theta})$ is always finite. Since we are interested in those $\boldsymbol{\rho}_{\mathcal{A}}$ for which the likelihood is infinitely large, we focus on the case where $\sigma^2 = 0$ and $\mathcal{A}^+(\boldsymbol{\delta}) \neq A^\diamond$ and partition the rest of the proof into three parts:

(a) For $\mathcal{A}^+(\boldsymbol{\delta}) \neq A^\diamond$ with cardinality

$$m_{\mathcal{A}^+} \;\triangleq\; \text{card}(\mathcal{A}^+(\boldsymbol{\delta})) \leq m_{A^\diamond} \tag{B.16a}$$

we have $\mathcal{A}^+(\boldsymbol{\delta}) \not\supseteq A^\diamond$ and $\text{card}(A^\diamond \cup \mathcal{A}^+(\boldsymbol{\delta})) \leq m_{\mathcal{A}^+} + m_{A^\diamond} < N$, see (3.5). Applying (B.1d) and (B.3c) for the index set $\mathcal{A}^+(\widehat{\boldsymbol{\delta}})$ (which satisfies the conditions of Lemma 2 (b)) yields

$$\lim_{\sigma^2 \searrow 0} \sigma^2 \ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \tfrac{1}{2} \lim_{\sigma^2 \searrow 0} \sigma^2 \ln(1/\sigma^2) \frac{\ln|P(\boldsymbol{\theta})|}{\ln(1/\sigma^2)} - \tfrac{1}{2} \lim_{\sigma^2 \searrow 0} \sigma^2\, \boldsymbol{y}^T\, P(\boldsymbol{\theta})\, \boldsymbol{y} < 0 \tag{B.16b}$$

and, consequently, $p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = 0$ at $\sigma^2 = 0$. The penalty is high for missing the (P_0)-optimal support.

(b) For $\mathcal{A}^+(\boldsymbol{\delta})$ with cardinality $m_{\mathcal{A}^+}$ that satisfies

$$m_{A^\diamond} < m_{\mathcal{A}^+} < N \tag{B.17a}$$

consider three cases: (i) $\mathcal{A}^+(\boldsymbol{\delta}) \not\supseteq A^\diamond$ and $\text{card}(A^\diamond \cup \mathcal{A}^+(\boldsymbol{\delta})) \leq N$, (ii) $\mathcal{A}^+(\boldsymbol{\delta}) \not\supseteq A^\diamond$ and $\text{card}(A^\diamond \cup \mathcal{A}^+(\boldsymbol{\delta})) > N$, and (iii) $\mathcal{A}^+(\boldsymbol{\delta}) \supset A^\diamond$, i.e. $\mathcal{A}^+(\boldsymbol{\delta})$ is strictly larger than $A^\diamond$. For (i), we apply the same approach as in (a) above, and conclude that $p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = 0$ at $\sigma^2 = 0$. For (ii), we observe that $\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) \leq -\frac{1}{2}\, N\, \ln(2\,\pi) + \frac{1}{2}\, \ln|P(\boldsymbol{\theta})|$ and apply (B.1d) for the index set $\mathcal{A}^+(\widehat{\boldsymbol{\delta}})$ (which satisfies the conditions of Lemma 2 (b)) to this upper bound, yielding

$$\lim_{\sigma^2 \searrow 0} \frac{-\frac{1}{2} N \ln(2\pi) + \frac{1}{2} \ln|P(\boldsymbol{\theta})|}{\ln(1/\sigma^2)} = \frac{N - m_{\mathcal{A}^+}}{2} < \frac{N - m_{A^\diamond}}{2}. \tag{B.17b}$$

Therefore, $\boldsymbol{\delta}$ that satisfy (ii) have 'smaller' likelihood (in the convergence speed sense defined in Lemma 4) than $\widehat{\boldsymbol{\delta}}$ in (B.14a) at $\sigma^2 = 0$. For (iii), arguments similar to those in (B.15) lead to

$$\lim_{\sigma^2 \searrow 0} \frac{\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta})}{\ln(1/\sigma^2)} = \frac{N - m_{\mathcal{A}^+}}{2} < \frac{N - m_{A^\diamond}}{2} \tag{B.18}$$

and, consequently, $\boldsymbol{\delta}$ that satisfy (iii) cannot match or outperform $\widehat{\boldsymbol{\delta}}$ at $\sigma^2 = 0$.

(c) For $\mathcal{A}^+(\boldsymbol{\delta})$ with cardinality $m_{\mathcal{A}^+} \geq N$, $P(\boldsymbol{\theta})$ is bounded and, therefore, $\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta})$ is finite. □

With a slight abuse of terminology, we refer to all $\boldsymbol{\rho}_{\mathcal{A}} = \widehat{\boldsymbol{\rho}}_{\mathcal{A}}$ defined by (B.14) as the ML estimates of $\boldsymbol{\rho}_{\mathcal{A}}$ under the scenario considered in Lemma 4. Interestingly, the proof of Lemma 4 reveals that, as $\sigma^2 \searrow 0$, $\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta})$ grows to

infinity when $\mathcal{A}^+(\boldsymbol{\delta}) \supset A^\diamond$ as well, but at a slower rate than that in (B.14c). In Corollary 5, we focus on the model where the index set A is equal to the (P_0)-optimal support $A^\diamond$ and, consequently, $B = B^\diamond = \mathcal{A}\backslash A^\diamond$.

Corollary 5: Assume that the URP and measurement number conditions (1) and (3) hold and consider the model with $A = A^\diamond$. Consider all variance-component estimates $\widehat{\boldsymbol{\rho}}_{A^\diamond} = \big(\widehat{\boldsymbol{\delta}}_{A^\diamond}, \widehat{\gamma}^2(A^\diamond), \widehat{\sigma}^2(A^\diamond)\big)$ that satisfy

$$\widehat{\boldsymbol{\delta}}_{A^\diamond} \succ \mathbf{0}_{m_A\times 1}, \quad \widehat{\gamma}^2(A^\diamond) = 0 \tag{B.19a}$$

$$\widehat{\sigma}^2(A^\diamond) = 0. \tag{B.19b}$$

Then,

$$\lim_{\sigma^2\searrow 0} \frac{\ln p(\boldsymbol{y}\,|\,\boldsymbol{\theta})}{\ln(1/\sigma^2)}\Big|_{A=A^\diamond,\boldsymbol{\delta}_{A^\diamond}=\widehat{\boldsymbol{\delta}}_{A^\diamond},\gamma^2=\widehat{\gamma}^2(A^\diamond)} = \tfrac{1}{2}\,(N - m_{A^\diamond}). \tag{B.20}$$

If $\sigma^2 > 0$, $p(\boldsymbol{y}\,|\,\boldsymbol{\theta})$ is always finite. Among all choices of $\boldsymbol{\delta}_{A^\diamond}$ and γ^2 for which $p(\boldsymbol{y}\,|\,\boldsymbol{\theta})$ is infinitely large, those $\boldsymbol{\delta}_{A^\diamond}$ and γ^2 defined by (B.19a) and (B.19b) 'maximize' the likelihood in the sense that $\ln p(\boldsymbol{y}\,|\,\boldsymbol{\theta})$ grows to infinity at the fastest rate as $\sigma^2 \searrow 0$, quantified by (B.20). Any choice of $\boldsymbol{\delta}_{A^\diamond}, \gamma^2$ different from $\widehat{\boldsymbol{\delta}}_{A^\diamond}, \widehat{\gamma}^2(A^\diamond)$ in (B.19a) *cannot* achieve this rate and, therefore, has a 'smaller' likelihood than $\widehat{\boldsymbol{\delta}}_{A^\diamond}, \widehat{\gamma}^2(A^\diamond)$ at $\sigma^2 = 0$.

Proof: Corollary 5 follows from the fact that the model $A = A^\diamond$ is nested within the full model $A = \mathcal{A}$. □

We refer to all $\widehat{\boldsymbol{\rho}}_{A^\diamond}$ defined by (B.19) as the ML estimates of $\boldsymbol{\rho}_{A^\diamond}$ under the scenario considered in Corollary 5.

Proof of Theorem 1: The conditions of Lemma 3 and Corollary 5 are satisfied, since they are included in the theorem's assumptions. Consider first the model $A = A^\diamond$; by Corollary 5, the ML variance-component estimates under this model are given in (B.19). Applying (B.20) and (B.7a) in Lemma 3 for $A = A^\diamond$ yields

$$\lim_{\sigma^2\searrow 0} \frac{\mathrm{GL}(\boldsymbol{\theta})}{\ln(1/\sigma^2)}\Big|_{A=A^\diamond,\boldsymbol{\delta}_A=\widehat{\boldsymbol{\delta}}_{A^\diamond},\gamma^2=\widehat{\gamma}^2(A^\diamond)} = \lim_{\sigma^2\searrow 0} \frac{\ln p(\boldsymbol{y}|\boldsymbol{\theta}) - \frac{1}{2}\ln|\mathcal{I}(\boldsymbol{\theta})|}{\ln(1/\sigma^2)}\Big|_{A=A^\diamond,\boldsymbol{\delta}_A=\widehat{\boldsymbol{\delta}}_{A^\diamond},\gamma^2=\widehat{\gamma}^2(A^\diamond)} = \tfrac{1}{2}\,(N - m_{A^\diamond} - 2). \tag{B.21}$$

Hence, under the conditions of Theorem 1, $\mathrm{GML}(A^\diamond)$ is infinitely large. In the following, we show that, for any other model $A \neq A^\diamond$, $\mathrm{GML}(A)$ in (3.1) is either finite or, if infinitely large, the rate of growth to infinity of $\mathrm{GL}(\boldsymbol{\theta})$ is smaller than that specified by (B.21). Actually, it suffices to demonstrate that any $\boldsymbol{\theta} = (A, \boldsymbol{\rho}_A)$ with $A \neq A^\diamond$ yields a 'smaller' $\mathrm{GL}(\boldsymbol{\theta})$ than $\boldsymbol{\theta} = (A^\diamond, \widehat{\boldsymbol{\rho}}_{A^\diamond})$, where $\widehat{\boldsymbol{\rho}}_{A^\diamond}$ has been defined in (B.19).

If $\sigma^2 > 0$, $P(\boldsymbol{\theta})$ is bounded and, therefore, the resulting $\mathrm{GL}(\boldsymbol{\theta})$ is always finite.

Consider the scenario where $\sigma^2 = 0$ and $\gamma^2 > 0$ and recall the definitions of $A^+(\boldsymbol{\delta}_A)$ and its cardinality m_{A^+} in (B.6). Then, $A^+(\boldsymbol{\delta}_A)\cup B$ is the set of indices corresponding to all positive signal variance components, with cardinality $m_{A^+} + m_B$. Now, consider two cases: (i) $m_{A^+} + m_B \geq N$ and (ii) $m_{A^+} + m_B < N$. For (i), the URP condition (1) implies that $P(\boldsymbol{\theta})$ is bounded and, therefore, $\mathrm{GL}(\boldsymbol{\theta})$ in (3.2) is finite. For (ii), observe that

$$\mathrm{GL}(\boldsymbol{\theta}) \leq -\tfrac{1}{2}N\ln(2\pi) + \tfrac{1}{2}\ln|P(\boldsymbol{\theta})| - \tfrac{1}{2}\ln|\mathcal{I}(\boldsymbol{\theta})| \tag{B.22}$$

apply (B.1d) in Lemma 1 *for the index set* $A^+(\boldsymbol{\delta}_A) \cup B$ (meaning that A and B in Lemma 1 have been replaced by $A^+(\boldsymbol{\delta}_A) \cup B$ and $\mathcal{A}\backslash[A^+(\boldsymbol{\delta}_A) \cup B]$, respectively), and use (B.7b) in Lemma 3 (b), yielding

$$\begin{aligned} \tfrac{1}{2} \lim_{\sigma^2 \searrow 0} \frac{-N \ln(2\,\pi) + \ln|P(\boldsymbol{\theta})| - \ln|\mathcal{I}(\boldsymbol{\theta})|}{\ln(1/\sigma^2)} &= \tfrac{1}{2}\,[N - (m_{A^+} + m_B) - 2\,(m_A - m_{A^+})] \\ &= \tfrac{1}{2}\,[N - m - (m_A - m_{A^+})] \leq \tfrac{1}{2}\,(N - m) < 0 \end{aligned} \tag{B.23}$$

where the last inequality follows from the assumption (2.2). Therefore, by (B.22), $\mathrm{GL}(\boldsymbol{\theta})$ goes to negative infinity as $\sigma^2 \searrow 0$. From (i)–(ii) above, we conclude that $\mathrm{GL}(\boldsymbol{\theta})$ cannot exceed $\mathrm{GML}(A^\diamond)$ when $\sigma^2 = 0$ and $\gamma^2 > 0$.

We now focus our attention to the scenario where $\sigma^2 = 0$ and $\gamma^2 = 0$. For any A and any corresponding $\boldsymbol{\delta}_A$, consider four cases: (i') $m_{A^+} \geq N$, (ii') $m_{A^+} \leq m_{A^\diamond}$ and $A^+(\boldsymbol{\delta}_A) \neq A^\diamond$, (iii') $m_{A^+} = m_{A^\diamond}$ and $A^+(\boldsymbol{\delta}_A) = A^\diamond$, and (iv') $m_{A^\diamond} < m_{A^+} < N$. For (i'), $P(\boldsymbol{\theta})$ is bounded and, therefore, $\mathrm{GL}(\boldsymbol{\theta})$ is finite. For (ii'), we have $\mathrm{card}(A^+(\boldsymbol{\delta}_A) \cup A^\diamond) \leq m_{A^+} + m_{A^\diamond} < N$ [see (3.5)] and, therefore,

$$\lim_{\sigma^2 \searrow 0} \sigma^2\,\mathrm{GL}(\boldsymbol{\theta}) = \tfrac{1}{2} \lim_{\sigma^2 \searrow 0} \Big[\sigma^2\,\ln(1/\sigma^2)\,\frac{\ln|P(\boldsymbol{\theta})|}{\ln(1/\sigma^2)} - \sigma^2\,\boldsymbol{y}^T\,P(\boldsymbol{\theta})\,\boldsymbol{y} - \sigma^2\,\ln(1/\sigma^2)\,\frac{\ln|\mathcal{I}(\boldsymbol{\theta})|}{\ln(1/\sigma^2)}\Big] < 0 \tag{B.24}$$

where we have applied (B.1d) in Lemma 1 and (B.3c) in Lemma 2 (b) for the index set $A^+(\boldsymbol{\delta}_A)$, and used (B.7a) in Lemma 3 (a); therefore, $\mathrm{GL}(\boldsymbol{\theta})$ goes to negative infinity as $\sigma^2 \searrow 0$. Here, Lemma 2 (b) delivers the severe penalty since $A^+(\boldsymbol{\delta}_A)$ does not include the (P_0)-optimal support $A^\diamond$.

If (iii') holds, we apply (B.1d) in Lemma 1 and (B.3a) in Lemma 2 (a) for the index set $A^+(\boldsymbol{\delta}_A) = A^\diamond$ and use (B.7a) in Lemma 3 (a), yielding

$$\lim_{\sigma^2 \searrow 0} \frac{\mathrm{GL}(\boldsymbol{\theta})}{\ln(1/\sigma^2)} = \tfrac{1}{2}\,[N - m_{A^\diamond} - 2\,(m_A - m_{A^\diamond} + 1)] \tag{B.25}$$

In this case, $m_A \geq m_{A^+} = m_{A^\diamond}$ and the largest possible (B.25) is attained if and only if $m_A = m_{A^+} = m_{A^\diamond}$, which is equivalent to $A = A^\diamond$; then, (B.25) reduces to (B.21). For $A \neq A^\diamond$, (B.25) is always smaller than the rate in (B.21), which is caused by inefficient modeling due to the zero variance components in the index set A; the penalty for this inefficiency is quantified by Lemma 3.

For (iv'), apply (B.1d) in Lemma 1 for the index set $A^+(\boldsymbol{\delta}_A)$ and use (B.7a) in Lemma 3 (a), yielding

$$\begin{aligned} \tfrac{1}{2} \lim_{\sigma^2 \searrow 0} \frac{-N \ln(2\,\pi) + \ln|P(\boldsymbol{\theta})| - \ln|\mathcal{I}(\boldsymbol{\theta})|}{\ln(1/\sigma^2)} &= \tfrac{1}{2}\,[N - m_{A^+} - 2\,(m_A - m_{A^+} + 1)] \\ &= \tfrac{1}{2}\,[N - m_{A^\diamond} - 2 - (m_A - m_{A^\diamond}) - (m_A - m_{A^+})] < \tfrac{1}{2}\,(N - m_{A^\diamond} - 2) \end{aligned} \tag{B.26}$$

where the inequality follows from $m_A \geq m_{A^+} > m_{A^\diamond}$; therefore, by (B.22), $\mathrm{GL}(\boldsymbol{\theta})$ cannot exceed $\mathrm{GML}(A^\diamond)$.

In summary, the model $A = A^\diamond$ maximizes $\mathrm{GML}(A)$ in (3.1) *globally and uniquely*. By (B.3b) in Lemma 2 (a),

$$\mathrm{E}_{\boldsymbol{s}\,|\,\boldsymbol{y},\boldsymbol{\theta}}[\boldsymbol{s}\,|\,\boldsymbol{y}, (A^\diamond, \widehat{\boldsymbol{\rho}}_{A^\diamond})] = \boldsymbol{s}^\diamond \tag{B.27}$$

where $\widehat{\boldsymbol{\rho}}_{A^\diamond} = (\widehat{\boldsymbol{\delta}}_{A^\diamond}, \widehat{\delta}^2_{B^\diamond}, \widehat{\sigma}^2(A^\diamond))$ is the set of ML variance-component estimates in (B.19) for $A = A^\diamond$. □

APPENDIX C
DERIVATION OF $\mathrm{GL}_{\mathrm{app}}(A, \boldsymbol{\delta}_A, \sigma^2)$

Plugging (4.14a) and (4.14c) into (2.6b) and applying (A.1a) yields

$$P(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\, I_N - \frac{1}{\sigma^2}\, H_A\, Z(\boldsymbol{\theta})\, H_A^T, \quad Z(\boldsymbol{\theta}) = [H_A^T\, H_A + \sigma^2\, D_A^{-1}(\boldsymbol{\delta}_A)]^{-1} \tag{C.1}$$

Approximating $H_A^T\, H_A$ by its diagonal elements (4.17), we have

$$\begin{aligned}
Z(\boldsymbol{\theta}) &\approx \mathrm{diag}\{z_{A,1}, z_{A,2}, \ldots, z_{A,m_A}\} && \text{(C.2a)}\\
I_{m_A} - H_A^T\, H_A\, Z(\boldsymbol{\theta}) &\approx \sigma^2\, \mathrm{diag}\{g_{A,1}, g_{A,2}, \ldots, g_{A,m_A}\} && \text{(C.2b)}\\
H_A^T\, H_A\, Z(\boldsymbol{\theta}) &\approx I_{m_A} - \sigma^2\, \mathrm{diag}\{g_{A,1}, g_{A,2}, \ldots, g_{A,m_A}\} && \text{(C.2c)}\\
\mathrm{tr}\{H_A^T\, H_A\, Z(\boldsymbol{\theta})\} &= \mathrm{tr}(I_{m_A}) - \mathrm{tr}[I_{m_A} - H_A^T\, H_A\, Z(\boldsymbol{\theta})] \approx m_A - \sigma^2 \sum_{i=1}^{m_A} g_{A,i} && \text{(C.2d)}
\end{aligned}$$

where

$$z_{A,i} = \frac{\delta_{A,i}^2}{\sigma^2 + h_{A,i}\, \delta_{A,i}^2}, \quad g_{A,i} = \frac{1 - h_{A,i}\, z_i}{\sigma^2} = \frac{1}{\sigma^2 + h_{A,i}\, \delta_{A,i}^2}, \quad i = 1, 2, \ldots, m_A \tag{C.2e}$$

and, to simplify notation, we have omitted the dependence of $z_{A,i}$ and $z_{A,i}$ on $\boldsymbol{\theta}$. Furthermore,

$$\begin{aligned}
\ln|P(\boldsymbol{\theta})| &= -N\, \ln(\sigma^2) + \ln|I_{m_A} - H_A^T\, H_A\, Z(\boldsymbol{\theta})| \approx -(N - m_A)\, \ln(\sigma^2) + \sum_{i=1}^{m_A} \ln g_{A,i} && \text{(C.2g)}\\
\mathrm{tr}[P^2(\boldsymbol{\theta})] &= \frac{N - m_A + \mathrm{tr}\{[I_{m_A} - H_A^T\, H_A\, Z(\boldsymbol{\theta})]^2\}}{(\sigma^2)^2} \approx \frac{N - m_A}{(\sigma^2)^2} + \sum_{i=1}^{m_A} g_{A,i}^2 && \text{(C.2h)}\\
H_A^T\, P(\boldsymbol{\theta})\, H_A &= \frac{H_A^T\, H_A - H_A^T\, H_A\, Z(\boldsymbol{\theta})\, H_A^T\, H_A}{\sigma^2} \approx \mathrm{diag}\{h_{A,1}\, g_{A,1}, \ldots, h_{A,m_A}\, g_{A,m_A}\} && \text{(C.2i)}\\
H_A^T\, P^2(\boldsymbol{\theta})\, H_A &= \frac{[I_{m_A} - H_A^T\, H_A\, Z(\boldsymbol{\theta})]\, H_A^T\, H_A\, [I_{m_A} - Z(\boldsymbol{\theta})\, H_A^T\, H_A]}{(\sigma^2)^2} \approx \mathrm{diag}\{h_{A,1}\, g_{A,1}^2, \ldots, h_{A,m_A}\, g_{A,m_A}^2\}. && \text{(C.2j)}
\end{aligned}$$

We approximate (3.4b)–(3.4d) using (C.2h)–(C.2j) and use $H_B\, H_B^T = I_N - H_A\, H_A^T$ [see (4.13) and (4.14b)]:

$$\begin{aligned}
\mathcal{I}_{\boldsymbol{\delta}_A, \boldsymbol{\delta}_A}(\boldsymbol{\theta}) &\approx \tfrac{1}{2}\, \mathrm{diag}\{h_{A,1}^2\, g_{A,1}^2, \ldots, h_{A,m_A}^2\, g_{A,m_A}^2\} && \text{(C.3a)}\\
\mathcal{I}_{\boldsymbol{\delta}_A, \gamma^2}(\boldsymbol{\rho}) &\approx \tfrac{1}{2}\, \left[h_{A,1}\, (1 - h_{A,1})\, g_{A,1}^2, \ldots, h_{A,m_A}\, (1 - h_{A,m_A})\, g_{A,m_A}^2\right]^T && \text{(C.3b)}\\
\mathcal{I}_{\gamma^2, \gamma^2}(\boldsymbol{\theta}) &\approx \frac{N - m_A}{2\, (\sigma^2)^2} + \tfrac{1}{2} \sum_{i=1}^{m_A} (1 - h_{A,i})^2\, g_{A,i}^2 && \text{(C.3c)}
\end{aligned}$$

yielding

$$\mathcal{I}_{\gamma^2, \gamma^2}(\boldsymbol{\theta}) - \mathcal{I}_{\boldsymbol{\delta}_A, \gamma^2}^T(\boldsymbol{\theta})\, \mathcal{I}_{\boldsymbol{\delta}_A, \boldsymbol{\delta}_A}^{-1}(\boldsymbol{\theta})\, \mathcal{I}_{\boldsymbol{\delta}_A, \gamma^2}(\boldsymbol{\theta}) \approx \frac{N - m_A}{2\, (\sigma^2)^2} \tag{C.3d}$$

and, using the formula for the determinant of a partitioned matrix [40, Th. 13.3.8]:

$$\ln|\mathcal{I}(\boldsymbol{\theta})| = \ln[\mathcal{I}_{\gamma^2, \gamma^2}(\boldsymbol{\theta}) - \mathcal{I}_{\boldsymbol{\delta}_A, \gamma^2}^T(\boldsymbol{\theta}) \mathcal{I}_{\boldsymbol{\delta}_A, \boldsymbol{\delta}_A}^{-1}(\boldsymbol{\theta}) \mathcal{I}_{\boldsymbol{\delta}_A, \gamma^2}(\boldsymbol{\theta})] + \ln|\mathcal{I}_{\boldsymbol{\delta}_A, \boldsymbol{\delta}_A}(\boldsymbol{\theta})| \approx \ln\left(\frac{N - m_A}{2\, (\sigma^2)^2}\right) + \sum_{i=1}^{m_A} \ln[\tfrac{1}{2} h_{A,i}^2\, g_{A,i}^2]. \tag{C.4}$$

Finally, the approximate GL formula (4.16) follows when we substitute (C.1), (C.2g), and (C.4) into (3.2)

REFERENCES

[1] I.F. Gorodnitsky and B.D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Processing*, vol. 45, pp. 600–616, Mar. 1997.

[2] E. Candès and J. Romberg, "Signal recovery from random projections," in *Computational Imaging III: Proc. SPIE-IS&T Electronic Imaging*, vol. 5674, C.A. Bouman and E.L. Miller (Eds.), San Jose, CA, Jan. 2005, pp. 76–86.

[3] E.J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, pp. 4203–4215, Dec. 2005.

[4] D. Malioutov, M. Çetin, and A.S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Processing*, vol. 53, pp. 3010–3022, Aug. 2005.

[5] E.J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate information," *Commun. Pure and Applied Mathematics*, vol. 59, pp. 1207–1233, Aug. 2006.

[6] *IEEE Signal Processing Mag. Special Issue on Sensing, Sampling, and Compression*, Mar. 2008.

[7] A.M. Bruckstein, D.L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, pp. 34–81, Mar. 2009.

[8] B.K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, 1995.

[9] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1998.

[10] E. Candès and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n," *Ann. Stat.*, vol. 35, pp. 2313–2351, Dec. 2007.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc., Ser. B*, vol. 58, pp. 267-288, 1996.

[12] M.A.T. Figueiredo, R.D. Nowak, and S.J. Wright, "Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems," *IEEE J. Select. Areas Signal Processing*, pp. 586–597, Dec. 2007.

[13] S. Mallat, Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[14] J. A. Tropp, A. C. Gilbert, " Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, pp. 4655–4666, Dec. 2007.

[15] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit (StOMP)", submitted for publication.

[16] D. Needell and J.A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comp. Harmonic Anal.*, vol. 26, pp. 301–321, May 2009.

[17] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inform. Theory*, vol. 55, pp. 2230–2249, May 2009.

[18] D.P. Wipf and B.D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Processing*, vol. 52, pp. 2153–2164, Aug. 2004.

[19] D.P. Wipf and B.D. Rao, "Comparing the effects of different weight distributions on finding sparse representations, " in *Advances in Neural Information Processing Systems*, Y. Weiss, B. Schölkopf and J. Platt (Eds.), Cambridge MA: MIT Press, vol. 18, 2006, pp. 1521–1528.

[20] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 56, pp. 2346–2356, Jun. 2008.

[21] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[22] K.K. Herrity, A.C. Gilbert, and J.A. Tropp, "Sparse approximation via iterative thresholding," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Toulouse, France, May 2006, pp. 624–627.

[23] T. Blumensath and M.E. Davies, "Iterative thresholding for sparse approximations," *J. Fourier Anal. Appl.*, vol. 14, pp. 629–654, Dec. 2008.

[24] T. Blumensath and M.E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comp. Harmonic Anal.*, vol. 27, pp. 265–274, Nov. 2009.

[25] T. Blumensath and M.E. Davies, "Normalised iterative hard thresholding; guaranteed stability and performance," to appear in *IEEE J. Select. Areas Signal Processing*, 2010.

[26] A. Maleki and D.L. Donoho, "Optimally tuned iterative thresholding algorithms for compressed sensing," to appear in *IEEE J. Select. Areas Signal Processing*, 2010.

[27] A. Dogandžić and K. Qiu, "Automatic hard thresholding for sparse signal reconstruction from NDE measurements," in *Proc. Annu. Rev. Progress Quantitative Nondestructive Evaluation*, Kingston, RI, Jul. 2009.

[28] B.D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Processing*, vol. 47, pp. 187–200, Jan. 1999.

[29] E.J. Candès, M.B. Wakin, and S.P. Boyd, "Estimating sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, pp. 877–905, Dec. 2008.

[30] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Las Vegas, NV, Apr. 2008, pp. 3869–3872.

[31] S.M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[32] M.H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *J. Amer. Stat. Assoc.*, vol. 96, pp. 746–774, Jun. 2001.

[33] J. Rissanen, *Information and Complexity in Statistical Modeling*. New York:Springer-Verlag, 2007.

[34] A. Dogandžić and K. Qiu, "ExCoV: Expansion-compression variance-component based sparse-signal reconstruction from noisy measurements," in *Proc. 43rd Annu. Conf. Inform. Sci. Syst.*, Baltimore, MD, Mar. 2009, pp. 186–191.

[35] S.M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[36] D.L. Donoho and J. Tanner "Counting faces of randomly projected polytopes when the projection radically lowers dimension," *J. Amer. Math. Soc.*, vol. 22, pp. 1–53, Jan. 2009.

[37] A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. R. Stat. Soc., Ser. B*, vol. 39, pp. 1–38, July 1977.
[38] G.J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, New York: Wiley, 1997.
[39] Å. Björck, *Numerical Methods for Least Squares Problems*. Philadelphia, PA: SIAM, 1996.
[40] D.A. Harville, *Matrix Algebra From a Statistician’s Perspective*. New York: Springer-Verlag, 1997.
[41] C. Luo, F. Wu, J. Sun, and C.W. Chen, “Compressive data gathering for large-scale wireless sensor networks,” in *Proc. Int. Conf. Mobile Comput. Networking (MobiCom)*, Beijing, China, Sept. 2009, pp. 145–156.