[Matthias W. Seeger and David P. Wipf]

Variational Bayesian Inference Techniques

[Improving and broadening the scope of compressive sensing]

© PHOTODISC

Milestones in sparse signal reconstruction and compressive sensing can be understood in a probabilistic Bayesian context, fusing underdetermined measurements with knowledge about low-level signal properties in the posterior distribution, which is maximized for point estimation. We review recent progress to advance beyond this setting. If the posterior is used as a distribution to be integrated over instead of merely an optimization criterion, sparse estimators with better properties may be obtained, and applications beyond point reconstruction from fixed data can be served. We describe novel variational relaxations of Bayesian integration, characterized as well as posterior maximization, which can be solved robustly for very large models by algorithms unifying convex reconstruction and Bayesian graphical model technology. They excel in difficult real-world imaging problems where posterior maximization performance is often unsatisfactory.

INTRODUCTION

Signal reconstruction from noisy measurements is a core problem in signal processing and computational mathematics. At its heart lies ambiguity resolution between alternative data explanations, based on uncertain knowledge about signal properties. A general approach is to model such knowledge probabilistically and then to invert this causal description for inference about the signal, given the data.

In this section, we phrase sparsity-penalized least squares reconstruction in a probabilistic Bayesian context, as maximization of the posterior distribution over signals conditioned on observed data. We motivate recent progress to advance beyond this setting, by embracing a different inference principle: Bayesian integration over the posterior, rather than its maximization. We review variational relaxations of Bayesian integration that not only result in estimators with provably better properties than posterior maximization, but also further applications beyond point reconstruction from fixed data. These relaxations are solved by convex reconstruction and Bayesian graphical model algorithms coming together, drawing a novel bridge between these concepts. In subsequent sections, we discuss large-scale algorithms, theoretical and empirical advancements, and demonstrate real-world improvements for magnetoencephalography (MEG) and electroencephalography (EEG) source localization and new applications to magnetic resonance imaging (MRI).

Digital Object Identifier 10.1109/MSP.2010.938082

1053-5888/10/$26.00©2010IEEE

SPARSE SIGNAL RECONSTRUCTION

Consider the linear reconstruction problem. Given measurements $y \in \mathbb{R}^m$ and design matrix $X \in \mathbb{R}^{m \times n}$, we seek $u \in \mathbb{R}^n$, which minimizes the squared error $\|y - Xu\|^2$. In MRI reconstruction, u is an image slice (n pixels), y are noisy Fourier coefficients, and X a partial discrete Fourier transform. With less measurements than pixels ($m < n$), this problem is ill posed: many different u give zero error. Ideally, estimation should be biased towards known properties of the signal class.

If we apply derivative or wavelet filters B to an image bitmap, the responses $s = Bu \in \mathbb{R}^q$ exhibit statistical sparsity: most values are tiny, however, some can be large [1]. We assume that $q \geq n$ in the sequel. An important special case is $B = I$. A remarkably robust low-level property of natural images, sparsity is what drives modern image compression and denoising methods. As sparsity of s is encouraged by way of the ℓ_p penalty $\|s\|_p^p = \sum_i |s_i|^p$ for $p \leq 1$ [2], the ℓ_p sparse reconstruction problem is biased towards images

$$\min_u \sigma^{-2}\|y - Xu\|^2 + 2\mathcal{R}_{\ell_p}(u), \quad \mathcal{R}_{\ell_p}(u) = \|Bu\|_p^p := \sum_{i=1}^{q} |s_i|^p,$$
$$p \in (0, 1],\ \sigma^2 > 0, s = Bu. \tag{1}$$

A particularly important instance is ℓ_1 reconstruction ($p = 1$), a convex optimization problem whose unique solution $\hat{u}_{\ell_1}$ is a tradeoff between data fit and signal sparsity. In general, $B\hat{u}_{\ell_p}$ is exactly sparse for $p \leq 1$, many of its coefficients are equal to zero. The strongest ℓ_0 penalty $\mathcal{R}_{\ell_0}(u) = \|s\|_0 := \sum_i \mathrm{I}_{\{s_i \neq 0\}}$ (which counts the number of nonzeros in s) leads to maximally sparse solutions, meaning a maximal number of elements equal to exactly zero.

With the advent of compressive sensing [2], [3], there has been growing interest in closely approximating maximally sparse reconstruction. However, problem (1) is nonconvex for any $p \in [0, 1)$, featuring many local minima. For p near zero, it becomes a combinatorial search, prohibitively expensive even for modest m, n, and q. For $B = I$, celebrated results establish that $\hat{u}_{\ell_1}$ has the same sparsity profile (location of nonzeros) as $\hat{u}_{\ell_0}$ whenever the design X satisfies a restricted isometry property (RIP) [2], [3]: roughly, each $2\|\hat{u}_{\ell_0}\|_0$ columns of X are close to orthonormal. While for randomly drawn X, RIPs hold with as little as $m = O(\|\hat{u}_{\ell_0}\|_0 \log n)$ measurements, they are not even remotely satisfied in many practical situations, where $\hat{u}_{\ell_1}$ tends to be much less sparse than $\hat{u}_{\ell_0}$ (see the section "Properties of Automatic Relevance Determination").

We can view ℓ_p reconstruction as a decision procedure based on a probabilistic sparse linear model (SLM). If $y = Xu + \varepsilon$, where ε is white Gaussian noise with variance σ^2, the data likelihood is $P(y|u) = N(Xu, \sigma^2 I)$. Since $-2\log P(y|u) = \sigma^{-2}\|y - Xu\|^2$ up to a constant, it matches the squared error term in (1), while the penalizer $\mathcal{R}(u)$ corresponds to a prior distribution $P(u)$ over signals: $\mathcal{R}(u) \propto -\log P(u)$. Statistical sparsity of $s = Bu$ is well captured by a Laplace prior distribution: $P(u) \propto \prod_i t_i(s_i)$ with

$$t_i(s_i) = e^{-\tau_i|s_i|}, \quad \tau_i > 0, \tag{2}$$

which corresponds to $\mathcal{R}_{\ell_1}(u)$ in (1). Another example is given by Student's *t* sparsity potentials

$$t_i(s_i) = (1 + (\tau_i/\nu)s_i^2)^{-(\nu+1)/2}, \quad \tau_i, \nu > 0, \tag{3}$$

where ν controls the degree of sparsity enforced. Combining $P(y|u)$ and $P(u)$ by rules of probability, we obtain the posterior distribution $P(u|y) \propto P(y|u)P(u)$, the general solution to our inference problem:

$$P(u|y) = Z^{-1}N(y|Xu, \sigma^2 I)\prod_{i=1}^{q} t_i(s_i), \quad s = Bu, \tag{4}$$

where $Z = \int N(y|Xu, \sigma^2 I)\prod_i t_i(s_i)\, du$ is known as the partition function. Bayesian inference amounts to computing posterior moments, such as the mean and (parts of the) covariance, which requires integration over (4). Sparse Bayesian inference is inference in SLMs.

The posterior is a distribution over signals, representing our uncertainty in what u should be. We can decide for a single point by maximum a posteriori (MAP) estimation: $\mathrm{argmax}_u P(u|y)$, or equivalently $\mathrm{argmin}_u - \log P(y|u) - \log P(u)$. Note that MAP estimation does not require integration over the posterior. For a Laplace sparsity prior (2) with $\tau_i = 1$, we recover ℓ_1 sparse reconstruction, and ℓ_p variants are obtained for $t_i(s_i) \propto e^{-\tau_i|s_i|^p}$. The Bayesian viewpoint provides a statistical context for linear reconstruction, within which a particular way of point reconstruction, MAP estimation, is equivalent to sparse reconstruction by penalized least squares (1).

SPARSITY PRIORS

Both statistical and computational properties of SLM inference methods are determined by the choice of positive potentials $t_i(s_i)$ in the prior $P(u) \propto \prod_i t_i(s_i)$. They allow us to enforce sparsity, a combinatorial property, within computationally tractable algorithms.

The statistical role of sparsity potentials is understood by inspecting the prior and posterior distributions they give rise to (Figure 1). For a high-dimensional Gaussian, which does not encourage sparsity, all coefficients s_i of typical samples are smallish, none are large or very small. In contrast, sparsity priors concentrate much more probability mass close to coordinate axes, and typical samples have many tiny and a few dominant $|s_i|$ [1], [4]. Conditioning on the same measurements, we obtain posterior distributions with markedly different properties [Figure 1(b)]. With sparsity priors, posterior mass is skewed towards coordinate axes, sparsity is enforced probabilistically, while Gaussian priors enforce nothing beyond uniformly small size. Note that sparsity priors have a distinct effect on the posterior mode: it is exactly sparse (see Figure 1(b), middle and right and the section "Sparse Signal Reconstruction"), a property not shared by its samples almost surely. For sparsity priors discussed in this article, posterior distributions concentrate mass on sparse points (thus promote sparsity exactly rather than statistically) only in limit cases, a notion we expand upon in the section "Benefits of Sparse Bayesian Inference."

Most sparsity potentials are super Gaussian [5]: they can be represented as maximum of Gaussian functions (see Figure 2 and the section "Variational Sparse Bayesian Inference"). Sparsity is enforced by non-Gaussian priors, yet their representations in terms of Gaussians allow for efficient algorithms. Among sparsity potentials, Laplacians stand out by being log-concave: $s_i \mapsto \log t_i(s_i)$ is concave. For such potentials, the posterior is unimodal with convex contours (Figure 1(b), left and middle), and MAP estimation (1) is a convex optimization problem. With non-log-concave priors, such as the Student's t (3), the posterior has multiple local modes in general (Figure 1(b), right). We will see in the section "Algorithms for Variational Sparse Bayesian Inference" that log-concavity can play much the same role for approximate Bayesian inference.

BENEFITS OF SPARSE BAYESIAN INFERENCE

Can sparse estimators with better properties than MAP estimation (1) be obtained from $P(\boldsymbol{u}|\boldsymbol{y})$? Moreover, sparse point reconstruction from given data being a means to an end, how can real-world applications be furthered by posterior information beyond its mode? In this section, we motivate advancements in sparse reconstruction and beyond, by using $P(\boldsymbol{u}|\boldsymbol{y})$ as a distribution to be integrated over, rather than a criterion to be maximized. Computational aspects are discussed in the section "Variational Sparse Bayesian Inference."

Shortcomings of MAP become evident for neuronal current source localization (see the section "Source Localization and Group Sparsity Penalization"), a typical real-world signal processing estimation problem. A smooth nonlinear model $\boldsymbol{f}(\cdot)$ is densely sampled at n locations $\boldsymbol{\theta}_i$, and sources are reconstructed from sensor readings $\boldsymbol{y}$ by sparse estimation with $X = [\boldsymbol{f}(\boldsymbol{\theta}_i)]$. Convex ℓ_1 reconstruction tends to perform poorly. Measurements are noisy and RIPs (see the section "Sparse Signal Reconstruction") are violated: columns of X are strongly correlated, a rule rather than an exception in real-world imaging applications. Nonconvex MAP reconstruction does not do well either: $-\log P(\boldsymbol{u}|\boldsymbol{y})$ has many shallow local minima, which efficient optimizers tend to get stuck in.

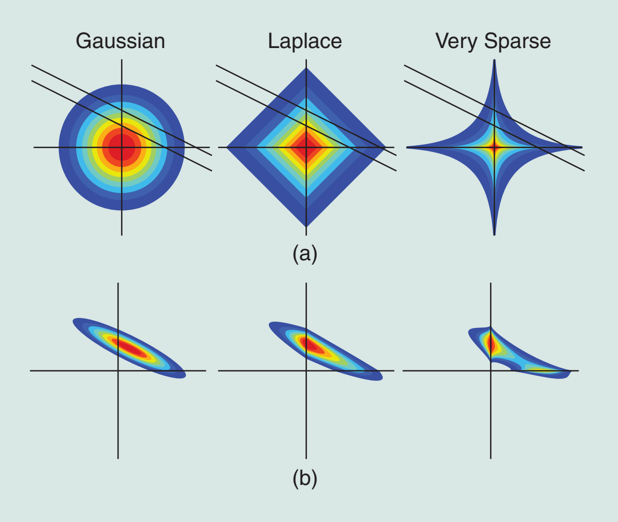
A Bayesian approach can alleviate these problems in many situations, computing the posterior mean $\mathrm{E}[\boldsymbol{u}\,|\,\boldsymbol{y}] = \int \boldsymbol{u}P(\boldsymbol{u}|\boldsymbol{y})\, d\boldsymbol{u}$ instead of its mode, integrating instead of maximizing over $P(\boldsymbol{u}|\boldsymbol{y})$. While the mean is not exactly sparse (see the section "Sparsity Priors"), this is enforced by taking a zero temperature limit, for example by computing $\mathrm{E}[\boldsymbol{u}|\boldsymbol{y}]$ for the Student's t potentials (3), then letting $\nu \to 0$. An approximation to this procedure, detailed in the section "Variational Sparse Bayesian Reconstruction," performs substantially better in source localization practice than ℓ_p sparse reconstruction for any $p \in (0, 1]$. The terminology "zero temperature limit" comes from statistical physics [6]. With decreasing temperature, the posterior concentrates on the "ground states," which are the ℓ_0 solutions in our case (see the section "Properties of Automatic Relevance Determination").

To motivate these advancements, note that $P(\boldsymbol{u}|\boldsymbol{y})$ is a probability density function, ranking $\boldsymbol{u}$ not by its height but by the mass surrounding it. For non-log-concave SLMs, mass tends to concentrate at deep optima, but many more shallow local optima stand for less sparse data explanations supported by very little posterior mass (Figure 1(b); right). MAP estimation is blind to mass and easily trapped in any shallow optimum, while such are mainly averaged out by Bayesian integration. Moreover, the posterior $P(\boldsymbol{u}|\boldsymbol{y})$ encodes dependencies between different signal hypotheses, which shape its moments more than its modes, and perfect reconstruction can be established under weaker conditions on X than RIPs for the ℓ_1 relaxation [7].

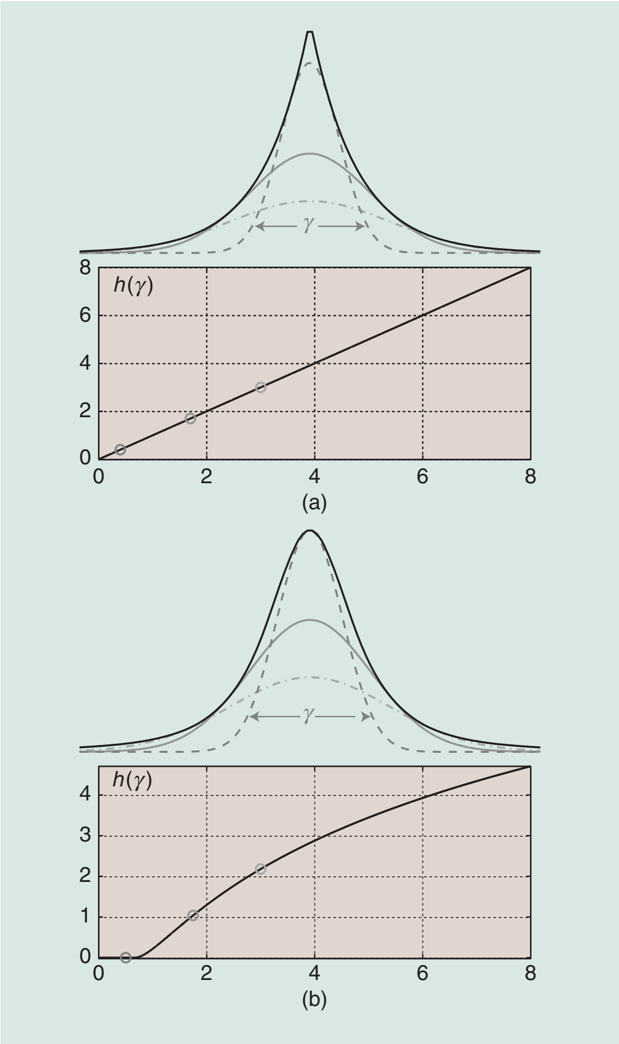
The posterior for a sparse linear model is useful far beyond point reconstruction. For example, prediction confidences are naturally provided by the posterior covariance matrix $\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}]$. Posterior covariance represents the structure of remaining uncertainty in $\boldsymbol{u}$, information different from any single best guess the like mode or mean, which allows to optimize data acquisition as such. Optimizing X by Bayesian experimental design can strongly improve subsequent sparse reconstruction [8], and is used to shorten scan time in MRI (see the section "Sampling Optimization of Magnetic Resonance Imaging").

VARIATIONAL SPARSE BAYESIAN INFERENCE

The advantages of Bayesian inference could well be offset by its computational difficulty. In general, given a high-dimensional function of the known structure, it can be much more difficult to accurately evaluate integrals over it than to find its mode.



[FIG1] (a) Different prior distributions with the same variance (Gaussian, Laplace (2), and $t(s) \propto e^{-\tau|s|^{0.4}}$), together with placement of one measurement (same for all). (b) Corresponding posteriors for same measurement. Mass is skewed towards the coordinate axes for sparsity priors, the mode is exactly sparse. Posteriors for Gaussian, Laplace are log-concave. (Figure courtesy of F. Steinke.)



[FIG2] Super-Gaussian potentials $t(s)$ admit tight Gaussian-form lower bounds of any width γ. Formally, $t(s) = \max_{\gamma \geq 0} e^{-s^2/(2\gamma) - h(\gamma)/2}$. (a) Laplace (2). (b) Student's t (3). (Figure courtesy of H. Nickisch.)

While there is a large and diverse body of approximate Bayesian inference technology, until recently none of these methods, applied to sparse linear models, could match the computational efficiency and theoretical characterization of MAP. In this section, we motivate a variational approximation to sparse Bayesian inference, which can be solved efficiently and reliably for very large SLMs and for which several advantages over MAP estimation can be established.

Bayesian inference in SLMs, integrating over the posterior (4), is intractable for two reasons coming together: $P(\boldsymbol{u}|\boldsymbol{y})$ is highly coupled (X is not block diagonal) and non-Gaussian. Two major classes of inference approximations are Markov chain Monte Carlo (MCMC) and variational relaxations [6]. In MCMC, $P(\boldsymbol{u}|\boldsymbol{y})$ is represented by samples, which are generated by random walks. While unbiased results are obtained in the infinite time limit, there are no realizable convergence diagnostics, making MCMC hard to use in practice. Moreover, standard MCMC samplers tend to converge very slowly for highly coupled models such as SLMs. While MCMC has recently been applied to sparse reconstruction [7], [9], it will play no further role in this article.

In variational approximations, Bayesian inference is relaxed to feasible optimization problems. While many different methods fall under this umbrella term (see [6] and [10] for a detailed overview), the particular approximation of interest here illustrates the main issues. Our goal is to fit $P(\boldsymbol{u}|\boldsymbol{y})$ by a Gaussian distribution $Q(\boldsymbol{u}|\boldsymbol{y};\boldsymbol{\gamma})$ parameterized by $\boldsymbol{\gamma}$, minimizing a divergence measure between $P(\boldsymbol{u}|\boldsymbol{y})$ and $Q(\boldsymbol{u}|\boldsymbol{y})$ (suppressing $\boldsymbol{\gamma}$-dependence for lighter notation), which we construct in the following way. We exploit super-Gaussianity of the prior potentials (see Figure 2 and the section "Sparsity Priors"), meaning that $t_i(s_i) = \max_{\gamma_i \geq 0} e^{-s_i^2/(2\gamma_i) - h_i(\gamma_i)/2}$ [5]. For example, for Laplace potentials (2), we have $h_i(\gamma_i) = \tau_i^2 \gamma_i$. Plugging these into the log partition function $\log Z = \log \int N(\boldsymbol{y}|X\boldsymbol{u}, \sigma^2 I) \prod_i t_i(s_i)\, d\boldsymbol{u}$ of the posterior (4), we obtain a representation purely in Gaussian terms. While still intractable, we note that the integral can easily be evaluated for any fixed $\boldsymbol{\gamma}$, as the log partition function of the Gaussian distribution $Q(\boldsymbol{u}|\boldsymbol{y}) \propto N(\boldsymbol{y}|X\boldsymbol{u}, \sigma^2 I) e^{-\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}/2}$. Each of these tractable Gaussian integrals lower bound $\log Z$, so that

$$\log Z \geq \max_{\boldsymbol{\gamma} \succeq 0} \log \int N(\boldsymbol{y}|X\boldsymbol{u}, \sigma^2 I) e^{-\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}/2 - h(\boldsymbol{\gamma})/2}\, d\boldsymbol{u},$$
$$\boldsymbol{s} = B\boldsymbol{u},\ \boldsymbol{\Gamma} := \operatorname{diag} \boldsymbol{\gamma}, \tag{5}$$

where $h(\boldsymbol{\gamma}) := \sum_{i=1}^q h_i(\gamma_i)$. The variational inference problem constitutes in optimizing this lower bound: we fit $Q(\boldsymbol{u}|\boldsymbol{y})$ to $P(\boldsymbol{u}|\boldsymbol{y})$ by maximizing the right-hand side of (5) or equivalently by minimizing the divergence criterion $-2\log \int N(\boldsymbol{y}|X\boldsymbol{u}, \sigma^2 I) e^{-\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}/2}\, d\boldsymbol{u} + h(\boldsymbol{\gamma})$. Note how both the approximation family $\{Q(\boldsymbol{u}|\boldsymbol{y})\}$ and divergence are implied by lower bounding $\log Z$ in a particular way. While this relaxation has been known for some time [5], [11], most properties discussed in this article are recent. Its application to SLMs is algorithmically and theoretically far better understood than for other approximations, and it can be solved for much larger models.

We can relate the variational inference problem (5) to MAP estimation directly: the latter is obtained from the former by replacing $\int \ldots d\boldsymbol{u}$ (integration over $\boldsymbol{u}$) with $\max_{\boldsymbol{u}}$ (optimization over $\boldsymbol{u}$). Advantages of variational Bayesian inference over MAP are ultimately due to this difference. A zero temperature limit case of our variational inference problem (the Student's t potentials, $\nu \to 0$; see the section "Benefits of Sparse Bayesian Inference") gives rise to a sparse point reconstruction method known as automatic relevance determination (ARD), (see the section "Variational Sparse Bayesian Reconstruction"), while instantiating (5) with proper priors (e.g., ℓ_p-based potentials with $p \in (0, 1]$ or the Student's t with $\nu > 0$) allows for Bayesian inference applications beyond sparse point estimation. The following points are discussed in the remainder of the article.

- For sparse reconstruction, ARD is an attractive alternative to convex or nonconvex MAP estimation (1). In the zero noise limit $\sigma^2 \to 0$, ARD's global minimum points are those of ℓ_0 reconstruction, yet it comes with far fewer local minimum points in general than strongly nonconvex MAP relaxations. Empirically, it can substantially outperform both ℓ_1 and nonconvex ℓ_p reconstruction in brain imaging applications, where columns of X are strongly correlated (see the section "Source Localization and Group Sparsity Penalization"). ARD, as opposed to MAP reconstruction, exploits nonuniform coefficient scaling of the ℓ_0 solution, a stable feature of real-world signals such as natural images (see the section "Properties of Automatic Relevance Determination").

- Variational sparse Bayesian inference (5) is a convex optimization problem if and only if MAP estimation is convex for the same model (see the section "Algorithms for Variational Sparse Bayesian Inference"). It is instrumental in driving nonlinear Bayesian experimental design, which can be used to optimize measurement designs X in real-world medical imaging settings (see the section "Sampling Optimization of Magnetic Resonance Imaging").
- The variational inference relaxation (5) is solved by double-loop algorithms, scaled up to very large models by reductions to convex reconstruction and Bayesian graphical model technology. Solving the ARD problem (locally) with these algorithms comes at the cost of a small number of reweighted ℓ_1 MAP problems (see the section "Properties of Automatic Relevance Determination").

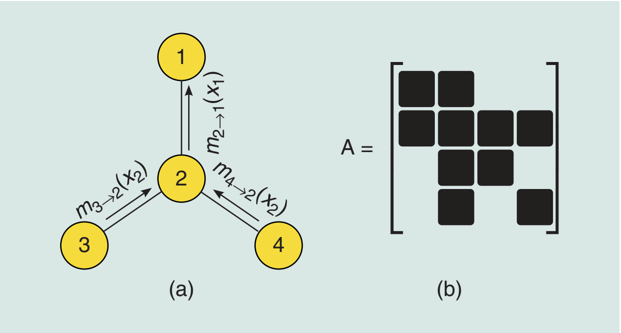
GAUSSIAN BAYESIAN GRAPHICAL MODELS

What does it take to solve the variational problem (5)? Can we use MAP estimation technology, or do we need computations of a different kind? At the least, we will need gradients w.r.t. $\boldsymbol{\gamma}^{-1}$

$$\nabla_{\boldsymbol{\gamma}^{-1}} \log \int N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2 \boldsymbol{I}) e^{-\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}/2} d\boldsymbol{u} = -(\mathrm{E}_Q[s_i|\boldsymbol{y}]^2 + \mathrm{Var}_Q[s_i|\boldsymbol{y}])/2.$$

We require means and variances of the marginal distributions $Q(s_i|\boldsymbol{y})$, Bayesian inference in Gaussian models. While common MAP algorithms (such as iteratively reweighted least squares) reduce to Gaussian mean computations equivalent to $(\mathrm{E}_Q[s_i|\boldsymbol{y}])$, variances are not part of them. Fortunately, we can approximate both means and variances by Bayesian graphical model algorithms.

An (undirected) graphical model describes structure in a family of multivariate probability distributions by way of an undirected graph $\mathcal{G}$, with nodes representing random variables (say, $x_1, \ldots, x_n$) and the absence of edges indicating conditional independence relationships. The latter can be used to dramatically simplify the computation of marginal posterior distributions $P(x_i)$, which by brute force scales exponentially in n. The formalism unifies ideas scattered across many disciplines: prominent examples are hidden Markov models, Kalman filtering, and low-density parity-check decoding [6]. Model distributions factorize into nonnegative potential functions defined on the cliques (fully connected node subsets) of $\mathcal{G}$: $P(x_1, \ldots, x_n) = Z^{-1} \prod_{C \in \mathcal{C}} \psi_C((x_i)_{i \in C})$, a representation in terms of local functions. In Figure 3, x_1, x_3, x_4 are separated by x_2, a structure that simplifies computations: $P(x_1) \propto \sum_{x_2, x_3, x_4} \Psi_{12}(x_1, x_2) \Psi_{23}(x_2, x_3) \Psi_{24}(x_2, x_4) = \sum_{x_2} \Psi_{12}(x_1, x_2) (\sum_{x_3} \Psi_{23}(x_2, x_3))(\sum_{x_4} \Psi_{24}(x_2, x_4))$. The basic elements of these computations are messages $m_{i \to j}(x_j)$ passed along edges of $\mathcal{G}$: $P(x_1) \propto m_{2 \to 1}(x_1)$ is obtained from $m_{3 \to 2}(x_2) \propto \sum_{x_3} \Psi_{23}(x_2, x_3)$ and $m_{4 \to 2}(x_2) \propto \sum_{x_4} \Psi_{24}(x_2, x_4)$ through the local sum-product message passing rule. For tree (singly connected) graphs $\mathcal{G}$, the message passing (or belief propagation) algorithm computes all marginals $P(x_i)$ with $2(n-1)$ sum-product operations [6].



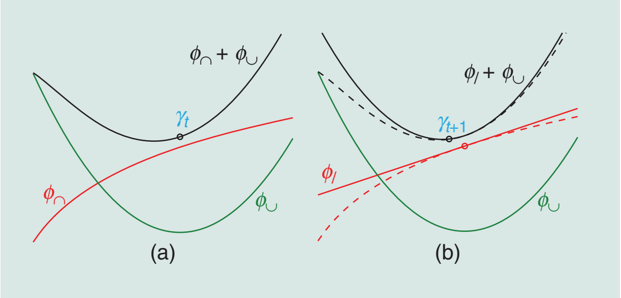
[FIG3] **(a) Tree-structured graphical model. Messages $m_{i\to j}(x_j)$, computed by the sum-product rule, are passed along edges. (b) Inverse covariance matrix structure for Gaussian model on the left (black squares indicating potential nonzero elements).**

For graphs with cycles, message passing becomes loopy belief propagation, an iterative algorithm for approximate inference, whose convergence and marginal error properties are subject to intense ongoing research [6]. The junction tree algorithm speeds up marginal computations even for loopy sparse $\mathcal{G}$, but remains generally intractable. Exact inference for discrete variable models is NP-hard [6]. For Gaussian models, exact inference requires $O(n^3)$ time and $O(n^2)$ space, which for very large n is practically prohibitive. $\mathcal{G}$ is determined by the sparsity pattern of the inverse covariance matrix $\boldsymbol{A} = \mathrm{Cov}[\boldsymbol{x}]^{-1}$: (ij) is an edge iff $a_{ij} \neq 0$ [Figure 3(b)], and the sum-product rule coincides with Kalman filtering equations. Whenever Gaussian loopy message passing converges (sufficient conditions for convergence are well understood), marginal means are correct, while variances are approximate in general [12]. Nevertheless, Gaussian message passing forms an integral part of today's most successful large scale variance approximations (see the section "Double-Loop Algorithms").

Gaussian inference fails to capture statistical sparsity (see the section "Sparsity Priors"), while sparse MAP estimation does not quantify prediction uncertainty and falls short of Bayesian averaging. However, we will see that their combination is sufficient to drive variational sparse Bayesian inference.

ALGORITHMS FOR VARIATIONAL SPARSE BAYESIAN INFERENCE

In this section, we describe efficient double-loop algorithms for solving the variational relaxation (5) at large scales and characterize its convexity. Full details are found in [13]–[15].



[FIG4] (a) Illustration of double-loop algorithm to minimize $\phi(\gamma) = \phi_\cap(\gamma) + \phi_\cup(\gamma)$, $\phi_\cap$ concave, $\phi_\cup$ convex. (b) Outer loop iteration starting at γ_t : (1) bound $\phi_\cap$ by affine $\phi_/$, tangent at γ_t, then (2) minimize convex upper bound $\phi_/(\gamma) + \phi_\cup(\gamma)$ to obtain γ_{t+1}. While the algorithm in the section "Double-Loop Algorithms" exploits concavity in γ^{-1}, the same principle applies. (Figure courtesy of H. Nickisch.)

Off-the-shelf optimization of (5) is not a viable option for very large models. Recall the form of the Gaussian posterior approximation $Q(u|y)$, parameterized by $\boldsymbol{\gamma} \succeq \mathbf{0}$. In particular, its covariance matrix is $\mathrm{Cov}_Q[u|y] = A^{-1}$, where $A := \sigma^{-2}X^TX + B^T\boldsymbol{\Gamma}^{-1}B$. As the integration is over a Gaussian, we can convert it into an optimization [13], obtaining the following reformulation of (5):

$$\min_{\boldsymbol{\gamma}\succeq 0}\min_{\boldsymbol{u}}\{\phi(\boldsymbol{u},\boldsymbol{\gamma}) := \log|A| + \sigma^{-2}\|\boldsymbol{y} - X\boldsymbol{u}\|^2 + \boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s} + h(\boldsymbol{\gamma})\}. \tag{6}$$

Recalling our comments in the section "Variational Sparse Bayesian Inference," the precise relationship between variational inference (5) and the MAP problem is clear now: their criteria differ solely in the $\log|A|$ term. Gaussian integration over $\boldsymbol{u}$ introduces dependencies between variables according to the posterior covariance (see the section "Benefits of Sparse Bayesian Inference"), giving rise to the coupling term $\log|A|$ not present in MAP. While the variational inference problem can be phrased as penalized least squares problem (1) with $\mathcal{R}_{\mathrm{VB}}(\boldsymbol{u}) = \min_{\boldsymbol{\gamma}\succeq 0}\log|A| + h(\boldsymbol{\gamma}) + \boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}$, this term does not come with the separable structure of MAP penalties (it cannot be expressed in the form $f(\boldsymbol{u}) = \sum_i f_i(s_i)$).

The reformulation (6) is essential for constructing efficient solvers. It is a jointly convex problem iff $h(\boldsymbol{\gamma})$ is convex, equivalent to all $t_i(s_i)$ being log-concave [14], [16]. Recalling the role of log-concavity for MAP (see the section "Sparsity Priors"), the variational inference problem is convex if and only if MAP estimation is convex for the same model. This property sets (5) apart from all other continuous-variable inference approximations we are aware of. Popular techniques like structured mean field [10] are nonconvex in general, others like expectation propagation [6] are not even provably convergent algorithms.

DOUBLE-LOOP ALGORITHMS

The joint minimization of (6) is difficult due to the coupled term $\log|A|$, but a concept known as concave-convex or majorize-minimize applies. The critical term is a concave function of $\boldsymbol{\gamma}^{-1}$. By Legendre duality, we have that $\log|A| = \min_{\boldsymbol{z}\succeq 0}\boldsymbol{z}^T(\boldsymbol{\gamma}^{-1}) - g_1^*(\boldsymbol{z})$ for some function $g_1^*(\boldsymbol{z})$, and as detailed in [13] and [14], (6) can be converted into

$$\min_{\boldsymbol{z}\succeq \boldsymbol{0}}\left(\min_{\boldsymbol{u}}\sigma^{-2}\|\boldsymbol{y} - X\boldsymbol{u}\|^2 - 2\sum_{i=1}^{q}\log t_i\sqrt{z_i + s_i^2} - g_1^*(\boldsymbol{z}).\right. \tag{7}$$

For any fixed $\boldsymbol{z}$, we have a separably penalized least squares problem w.r.t. $\boldsymbol{u}$ of the same form (1) as MAP estimation (in fact, MAP estimation would be obtained precisely by setting all $z_i = 0$). For Laplace potentials (2), this inner problem is (1) with $\mathcal{R}(\boldsymbol{u}) = \sum_i \tau_i\sqrt{z_i + s_i^2}$.

Our double-loop algorithm [13], [15] iterates between inner loop minimizations of (7) over $\boldsymbol{u}$ (which involve posterior mean calculations ($\mathrm{E}_Q[s_i|\boldsymbol{y}]$) as commonly used for MAP estimation), and outer loop updates of $\boldsymbol{z}$ (see Figure 4). The latter are given by $\boldsymbol{z} \leftarrow \nabla_{\boldsymbol{\gamma}^{-1}}\log|A| = \mathrm{diag}^{-1}(BA^{-1}B^T) = (\mathrm{Var}_Q[s_i|\boldsymbol{y}])$, they require computing Gaussian variances. The variational relaxation (5) is solved at large scales by penalized least squares reconstruction and Gaussian model inference joining forces. The algorithm is guaranteed to converge to a stationary point [15], whether (5) is convex or not. It converges orders of magnitude faster on large SLMs than other approximate inference methods we are aware of. For non-log-concave prior potentials, a simple variant of (7) ensures that merely convex inner loop problems have to be solved [16].

While in Gaussian models, posterior means ($\mathrm{E}_Q[s_i|\boldsymbol{y}]$) are obtained solving a single linear system by the conjugate gradients algorithm, no similarly general and efficient method is known for the variances ($\mathrm{Var}_Q[s_i|\boldsymbol{y}]$). Today's most promising variance approximations use graphical model message passing (see the section "Gaussian Bayesian Graphical Models") or fit A by a low rank matrix [17], for example, by using the Lanczos algorithm [13], [18]. Message passing is used as subroutine in methods that approximate $Q(\boldsymbol{u}|\boldsymbol{y})$ by tree-like graphs [17], [19] or in Gauss-Seidel algorithms [20]. Distributed message passing computations can be used together with Lanczos methods to address very large models (see the section "Sampling Optimization of Magnetic Resonance Imaging"). Effects of low rank variance approximation errors on the problem (7) are analyzed in [21].

VARIATIONAL SPARSE BAYESIAN RECONSTRUCTION

Bayesian inference can be used for sparse point reconstruction by computing the posterior mean $\int \boldsymbol{u}P(\boldsymbol{u}|\boldsymbol{y})\,d\boldsymbol{u}$ in a zero temperature limit (see the section "Benefits of Sparse Bayesian Inference"), where posterior mass is concentrated on exactly sparse points. A variational approximation thereof, known as ARD [15], is obtained from (5) with Student's t potentials (3), letting $\nu \to 0$, which renders $h(\boldsymbol{\gamma}) = \sum_i \log\gamma_i$ [14]. The role of this concave function, unbounded below as $\gamma_i \to 0$, is to drive components of $\boldsymbol{\gamma}$ to zero. It is easy to see that $\{\gamma_i = 0\}$ implies the elimination of s_i: the Gaussian $Q(\boldsymbol{u}|\boldsymbol{y})$ is a degenerate distribution, fixing components of $\boldsymbol{s}$ to zero. In contrast, degenerate $Q(\boldsymbol{u}|\boldsymbol{y})$ cannot arise in variational inference for normalizable potentials $t_i(s_i)$, where $\boldsymbol{\gamma} > 0$ throughout.

REWEIGHTED ℓ_1 ALGORITHM

In the ARD zero temperature limit, we can use an alternative to the double-loop algorithm above, enjoying the same global convergence property but some additional benefits [15]. Since $\boldsymbol{\gamma} \mapsto \log|A| + \sum_i \log\gamma_i$ is concave for $\boldsymbol{\gamma} \succeq \mathbf{0}$ [14], Legendre duality provides its representation as $\min_{z \succeq 0} \boldsymbol{z}^T\boldsymbol{\gamma} - g_2^*(\boldsymbol{z})$ (see Figure 4) $g_2^*(\boldsymbol{z})$, and as detailed in [15], (6) is converted into

$$\min_{z \succeq 0}\left(\min_{u} \sigma^{-2}\|\boldsymbol{y} - X\boldsymbol{u}\|^2 + 2\sum_{i=1}^{q} z_i^{1/2}|s_i|\right) - g_2^*(\boldsymbol{z}). \tag{8}$$

In this case, the inner loop problem is a reweighted form of ℓ_1 reconstruction (1), whose minimizer is exactly sparse. Running this double-loop algorithm often requires many less outer loop iterations than the method of the "Double-Loop Algorithms" section applied to ARD, since the bound on $\log|A|$ used here is tighter for $\gamma_i \approx 0$. Moreover, it is easier to add additional convex constraints on $\boldsymbol{u}$ [22]. Outer loop updates are given by $z_i \leftarrow (1 - \mathrm{Var}_Q[s_i|\boldsymbol{y}]/\gamma_i)/\gamma_i$. If $\|\boldsymbol{\gamma}\|_0 \ll q$, these values can be computed efficiently by low-rank formulae [22], [23].

PROPERTIES OF AUTOMATIC RELEVANCE DETERMINATION

In this section, we show that ARD can offer substantial advantages over separable (convex or nonconvex) MAP estimation when searching for maximally sparse solutions. These improvements have to be offset against an increase in running time, since ℓ_1 reconstruction has to be run a few times (see the section "Reweighted ℓ_1 Algorithm.") Detailed accounts of these results, including all proofs, are found in [22] and [23]. We will assume that $m < n$ (less observations than signal components), and that each subset of m columns of X is linearly independent. Moreover, $B = I$ in this section, and $\boldsymbol{s} = \boldsymbol{u}$. For simplicity, we restrict ourselves to the zero noise limit ($\sigma^2 \to 0$), for which ℓ_0 reconstruction becomes

$$\min_{u}\left\{\|\boldsymbol{u}\|_0 = \sum_{i=1}^{n} \mathrm{I}_{\{u_i \neq 0\}}\right\} \quad \text{such that } \boldsymbol{y} = X\boldsymbol{u}, \tag{9}$$

while ARD becomes

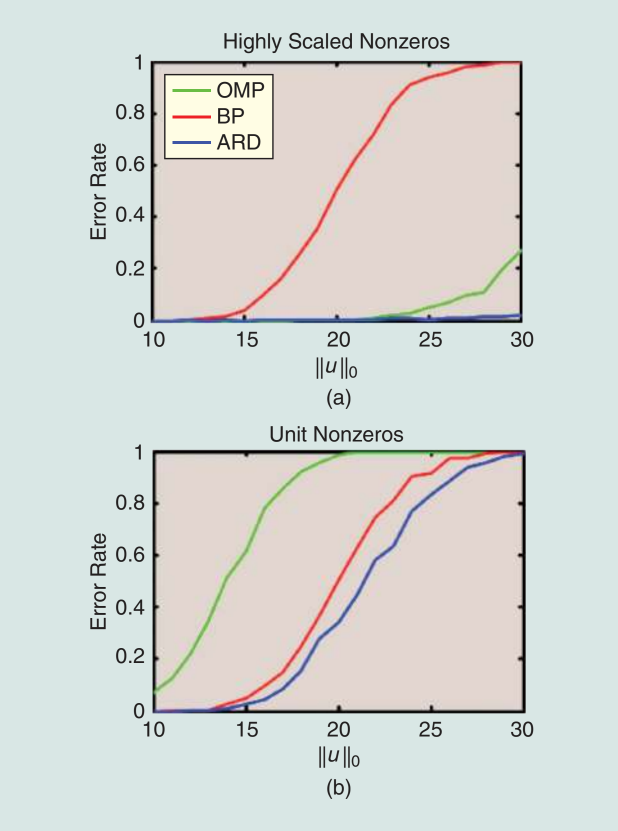
$$\min_{u} \mathcal{R}_{\mathrm{VB}}(\boldsymbol{u}) \quad \text{such that } \boldsymbol{y} = X\boldsymbol{u}, \tag{10}$$

where

$$\mathcal{R}_{\mathrm{VB}}(\boldsymbol{u}) = \min_{\gamma \succeq 0} \log|X\boldsymbol{\Gamma}X^T| + \boldsymbol{u}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{u} = \min_{z \succeq 0} 2\sum_{i=1}^{n} z_i^{1/2}|u_i| - g_2^*(\boldsymbol{z})$$

(which holds up to an additive constant [14]).

Since the ℓ_1 MAP relaxation of maximally sparse ℓ_0 reconstruction is exact when RIPs hold true (see the section "Sparse Signal Reconstruction"), we focus on practically relevant situations where such conditions for X are violated (see the section "Source Localization and Group Sparsity Penalization"). There is growing empirical evidence that the variational ARD method can substantially outperform separable MAP-like relaxations in many such cases (for example, Figure 5), and our aim is to provide sound explanations for these findings. Our results do not necessarily imply that ARD improves upon MAP reconstruction uniformly over all sufficiently sparse instances, but rather that it exploits additional structure in the signal beyond exact sparsity, thus is in general much less reliant on X obeying RIPs to work well.



[FIG5] Comparison of ARD sparse Bayesian reconstruction (10), using the algorithm from the section "Double-Loop Algorithms," with basis pursuit (BP), solving the ℓ_1 problem (1) for $\sigma^2 = 0$, and orthogonal matching pursuit (OMP), a greedy method for solving (9) locally. Data generation: $m = 50$, $n = 100$, columns of X drawn uniformly of unit norm, true vectors $\boldsymbol{u}'$ with support size $d \in [10, 30]$, nonzeros either from (a) highly scaled distribution or (b) set to one, $\boldsymbol{y} = X\boldsymbol{u}'$ (m, n, d guarantee that $\boldsymbol{u}'$ maximally sparse in general). Error rates (fraction of failure to find sparsity pattern of $\boldsymbol{u}'$) were estimated by running 1,000 repeats each. (a) As expected, BP performs identical in both settings, while OMP and especially ARD benefit from nonuniform scaling. (b) Even in the worst case of identical nonzeros, ARD outperforms the other methods.

Separable MAP-like relaxations often fail to closely approximate (9) for one of the following reasons. For a convex MAP relaxation (e.g., ℓ_1 reconstruction), the global solution is not sufficiently sparse, thus biased away from maximally sparse solutions of (9). One way to decrease such bias is to employ MAP with stronger, non-log-concave potentials (e.g., minimizing ℓ_p with $p \approx 0$, or running MAP with the Student's t in the $\nu \to 0$ limit). However, such criteria are notoriously hard to optimize, and many algorithms get stuck in poor local solutions. Another option is to employ a nonseparable penalizer $\mathcal{R}(\boldsymbol{u})$. This is the route taken by ARD, whereby the penalizer $\mathcal{R}_{\mathrm{VB}}(\boldsymbol{u})$ is obtained in a principled manner through Bayesian integration. In the zero noise limit, ARD (10) has precisely the same global maximum points as the ℓ_0 problem (9). However, while nonconvex MAP relaxations sharing this global minima condition are plagued by a provably large number of local minima, Bayesian averaging serves to smooth away many (and typically most) suboptimal local solutions in the variational ARD criterion.

It is understood in great detail by now how the difficulty of sparse reconstruction grows with the number of nonzeros in the optimal ℓ_0 solution $\hat{\boldsymbol{u}}_{\ell_0}$. However, the size distribution of nonzeros in $\hat{\boldsymbol{u}}_{\ell_0}$ can play a substantial role as well. This fact has largely been overlooked so far [2], quite possibly because the recovery performance of the ℓ_1 MAP relaxation is invariant to rescaling the nonzeros in $\hat{\boldsymbol{u}}_{\ell_0}$. Nonuniform coefficient scaling is a property of most real-world signal classes of interest (for example, distributions of natural image wavelet coefficients are scale-free and vary over a large dynamic range [1]). In contrast to ℓ_1 reconstruction, such scaling is successfully exploited by ARD whenever present (see Figure 5). The following result confirms these observations: whenever the coefficient scaling of $\hat{\boldsymbol{u}}_{\ell_0}$ is sufficiently nonuniform, the ℓ_0 solution is the only local minimum point of the ARD criterion. Say that $\boldsymbol{u}'$ with $\|\boldsymbol{u}'\|_0 \leq k$ obeys scaling constraints $(\varphi_i) \in (0, 1]^{k-1}$ of order k if $|u'_{(i+1)}| \leq \varphi_i |u'_{(i)}|$ for $i = 1, \ldots, k-1$, where $(u'_{(i)})$ is a permutation of $\boldsymbol{u}'$ in nonincreasing (absolute) coefficient ordering: $|u'_{(i+1)}| \geq |u'_{(i)}|$. Then, for any X, there exists scaling constraints of order $m-2$, so that for any signal $\boldsymbol{u}'$, $\|\boldsymbol{u}'\|_0 < m$ obeying these constraints and $\boldsymbol{y} = X\boldsymbol{u}'$, ARD has *no* local minimum point apart from $\boldsymbol{u}'$, and $\boldsymbol{u}'$ will necessarily equal the unique, maximally sparse solution.

Finally, as the reweighted ℓ_1 double-loop algorithm from the section "Reweighted ℓ_1 Algorithm" is typically started with straight ℓ_1 reconstruction ($\boldsymbol{z} = \mathbf{1}$ for the first outer loop step), it is important to stress that apart from increased running time, there is no risk involved in running further iterations and progressing beyond convex MAP estimation. The sparsity $\|\boldsymbol{u}_*\|_0$ for successive inner loop minimizers $\boldsymbol{u}_*$ is nonincreasing, so that ARD's recovery performance cannot be worse than that of ℓ_1 MAP estimation. Moreover, given any X and sparsity profile $\mathcal{S} \subset \{1, \ldots, n\}$ [location of nonzeros; $|\mathcal{S}| \leq (m+1)/2$] for which ℓ_1 reconstruction fails to recover some $\hat{\boldsymbol{u}}_{\ell_0}$, running subsequent ARD iterations can always lead to successful recovery. Precisely, there are sets $\{\boldsymbol{y} = X\boldsymbol{u}'\}$ of nonzero Lebesgue measure, $\boldsymbol{u}'$ with sparsity profile $\mathcal{S}$, for which the ARD reweighted ℓ_1 algorithm always succeeds in solving (9), yet ℓ_1 reconstruction (1), the tightest convex relaxation of the ℓ_0 problem, always fails. This statement in particular covers instances for which RIPs do not hold.

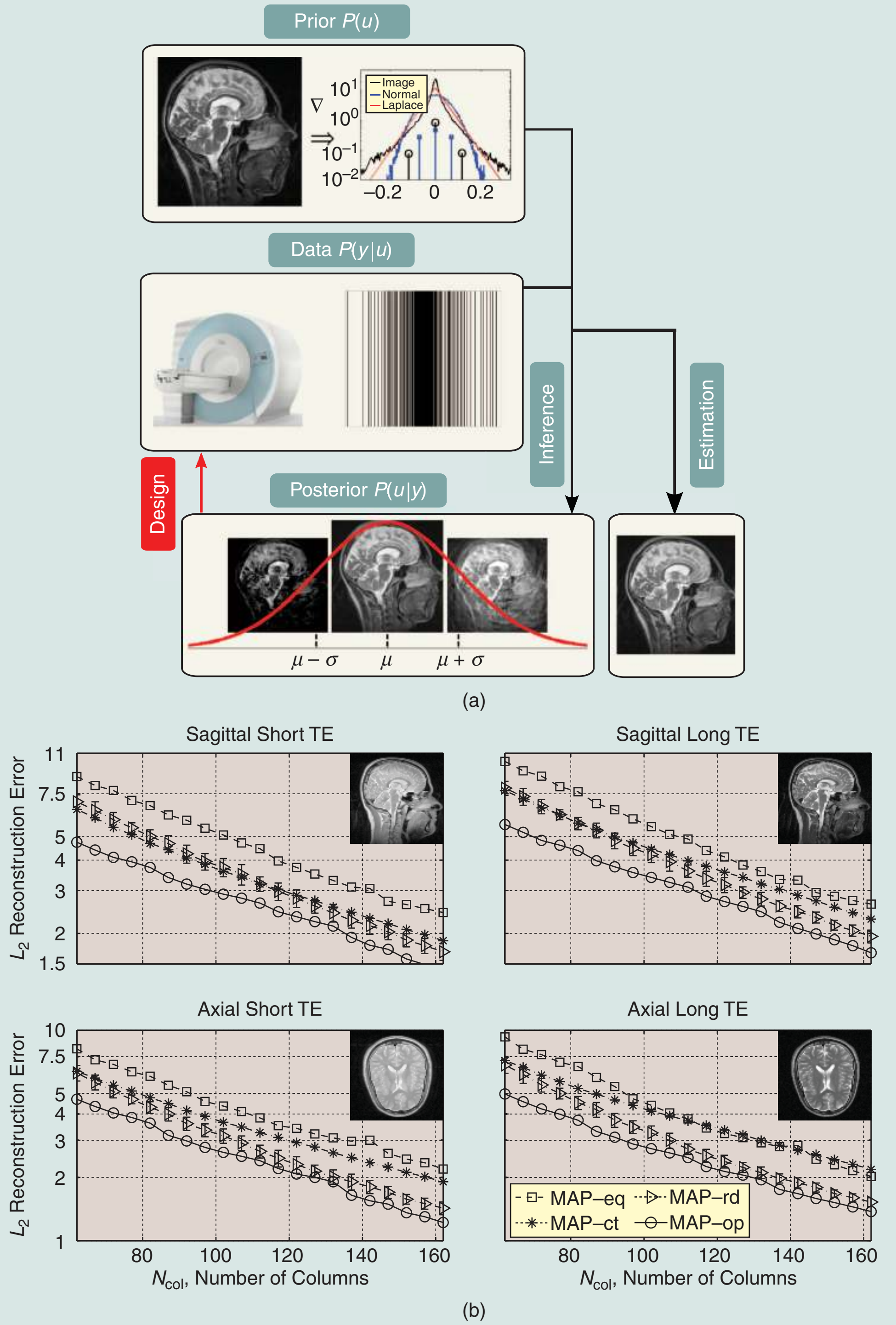
APPLICATIONS

Bayesian methods for sparse linear models are useful for sparse point reconstruction, and beyond for decision making based on uncertain, highly underdetermined knowledge. In this section, we provide examples for sparse Bayesian inference and point reconstruction of particular interest to signal and image processing.

SAMPLING OPTIMIZATION OF MAGNETIC RESONANCE IMAGING

In MRI [24], image slices are reconstructed from coefficients sampled along smooth trajectories in Fourier space (phase encodes). In Cartesian MRI, phase encodes are dense columns (or rows) in discrete Fourier space. The most serious limiting factor is long scan time. MRI is a prime candidate for compressive sensing in practice [25], [26]: if quality images can be reconstructed from an undersampling, time is saved at no additional costs. The success of sparse reconstruction on realistic images is mainly determined by the choice of the design X [8], [26]. This empirically well-established fact, not captured by current compressive sensing theory, motivates the optimization of X, which can be done with Bayesian experimental design. A good explanation for the apparent mismatch between theory and real imaging practice is that assumptions made by the theory do not come close to capturing real-world image structure [8].

Bayesian experimental design makes use of posterior covariance $\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}]$, quantifying the dependency structure of remaining uncertainty in $\boldsymbol{u}$, in that subsequent phase encodes are aligned with directions of maximum uncertainty, optimizing X in a greedy sequential manner. Our algorithm iterates between scoring phase encodes X_* based on the posterior approximation $Q(\boldsymbol{u}|\boldsymbol{y})$ for the current data $(X, \boldsymbol{y})$, extending X by the winner X_* and $\boldsymbol{y}$ by corresponding new data $\boldsymbol{y}_*$, and refitting $Q(\boldsymbol{u}|\boldsymbol{y})$ [Figure 6(a)]. The scoring criterion to be maximized is $I_{Q(\boldsymbol{u}|\boldsymbol{y})}(X_*) = (1/2)\log|\boldsymbol{I} + X_*^T A^{-1} X_*|$, an approximation to the information gain [27], [4]. Computing $\{I_{Q(\boldsymbol{u}|\boldsymbol{y})}(X_*)\}$ for a large set of candidate encodes X_* is closely related to computing Gaussian variances $(\mathrm{Var}_Q[s_j|\boldsymbol{y}])$ for all components s_j, and essentially the same technology can be applied (see the sections "Gaussian Bayesian Graphical Models" and "Double-Loop Algorithms"). This procedure was implemented for a study with Cartesian MRI scans of the human brain [13], [26], driven by the convex variational relaxation (5) with Laplace prior potentials (here, $n = 131{,}072$, $q \approx 3n$, m up to $(3/4)n$, and $\boldsymbol{u}$ are complex-valued). Optimized designs X clearly outperform setups drawn at random from engineered variable-density ensembles [25], when either are applied to a wide variety of test data (Figure 6(b); note that all reconstructions for given designs X are done by conventional ℓ_1 MAP estimation). This framework is not specific to MRI and applies to other image acquisition modalities as well. It is an instance of adaptive compressive sensing, choosing X actively based on representative real-world data using concepts from machine learning. Adaptive compressive sensing schemes have to maintain some representation of uncertainty over all coefficients beyond single best estimates, such as the measurement budget distribution



[FIG6] (a) Bayesian experimental design for speeding up MRI acquisition. (b) Comparison of ℓ_1 MAP reconstruction for designs X chosen in different ways: (op) optimized by Bayesian method; (rd) sampled from variable-density ensemble [25]; (ct) dense low frequencies; (eq) equi-spaced columns. Shown is ℓ_2 distance $\| |u^*| - |u_{\text{true}}| \|$ (u_{true} reconstruction from Nyquist-dense measurements) versus size of X (number of columns in Fourier space), on wide variety of data not used for design optimization. (Figure used with permission from [26].)

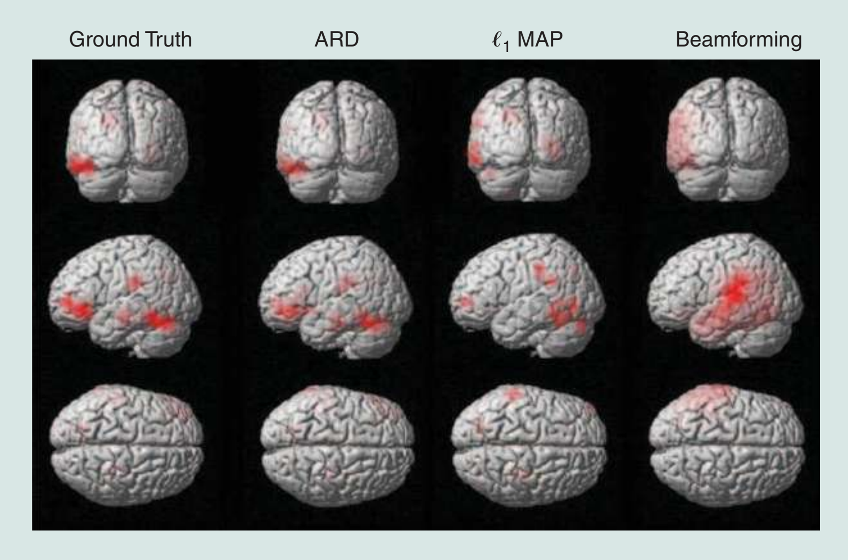
in distilled sensing [28], a recent adaptive reconstruction method. Distilled sensing is an online adaptive scheme (X has to be re-learned for each reconstruction), while Bayesian experimental design aims to find designs that generalize well to unseen data. Moreover, distilled sensing has not been applied to real-world images.

In realistic MRI experiments, a stack of neighboring image slices is acquired in an interleaved fashion. Bayesian design optimization can be generalized to this setting, representing dependencies between slices, if the double-loop algorithm is configured with parallel convex reconstruction and approximate Kalman smoothing [29]. The Markov structure of this very large setup is exploited by graphical model technology to iteratively reduce computations to the single slice case (see the section "Gaussian Bayesian Graphical Models").

SOURCE LOCALIZATION AND GROUP SPARSITY PENALIZATION

A basic problem in array processing is source localization [30]. Measurements $\boldsymbol{y} \in \mathbb{R}^m$ are modeled as $P(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\Theta}) = N\left(\boldsymbol{y}|\sum_{j=1}^{k}\alpha_j f(\boldsymbol{\theta}^{(j)}), \sigma^2\boldsymbol{I}\right)$, where $\boldsymbol{\Theta} = [\boldsymbol{\theta}^{(j)}] \in \mathbb{R}^{r\times k}$ represents k source locations, $\boldsymbol{\alpha} \in \mathbb{R}^k$ signal amplitudes, and $f(\cdot) : \mathbb{R}^r \to \mathbb{R}^m$ is a fixed nonlinear signal transduction function. The number of active sources k must be learned along with their locations. Estimating $\boldsymbol{\alpha}$, $\boldsymbol{\Theta}$ and k is an intricate nonconvex optimization problem, a powerful alternative to which is offered by sparse estimation. We densely sample locations at $n \gg m$ points $\{\boldsymbol{\theta}_i\}$ and apply sparse Bayesian reconstruction with $X = [f(\boldsymbol{\theta}_i)]$. Upon convergence to $\boldsymbol{u}_*$, the nonzeros correspond to $\boldsymbol{\alpha}$, $k \leftarrow \|\boldsymbol{u}_*\|_0$, and relevant locations $\boldsymbol{\theta}^{(j)}$ correspond to active columns of X. Due to favorable properties described in the section "Properties of Automatic Relevance Determination," estimates are much less dependent on initialization or the local minimum profile of the likelihood than in the traditional nonlinear setup. Sparse Bayesian source localization has been applied successfully to tomographic imaging of neuronal current sources by MEG and EEG [31]. The ARD double-loop algorithms discussed above scale well to realistic problem sizes, e.g., $m = 300$ sensors and $n = 10^6$ voxels, significantly outperforming results for ℓ_1 MAP reconstruction (see Figure 7). As noted in the section "Benefits of Sparse Bayesian Inference," this can be explained by RIPs [3] certainly being violated: due to dense sampling and smoothness of $f(\cdot)$, columns of X are strongly correlated. Similar to MRI (see the section "Sampling Optimization of Magnetic Resonance Imaging") or most imaging modalities in practice, this property of X is not negotiable. Other successful applications of ARD include direction-of-arrival estimation for sonar and radar processing [30], [32].

Our algorithms are easily extended to incorporate group sparsity penalization [33], [34]. If $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$ decomposes into subvectors $\boldsymbol{s}_i$ (for example, columns of a matrix), we may replace $t_i(s_i)$ (scalar s_i) by $t_i(\|\boldsymbol{s}_i\|)$, $\|\cdot\|$ the Euclidean norm. For example, if $\boldsymbol{s}$ is complex-valued (see the section "Sampling Optimization of Magnetic Resonance Imaging"), the encoding $\boldsymbol{s}_i \in \mathbb{R}^2$ naturally leads to group penalization. A more general example is the simultaneous sparse approximation problem, arising in applications such as image coding or source localization [30]: given an overcomplete dictionary and a set of response vectors $\{\boldsymbol{y}_k\}$, the goal is to jointly encode them using the same sparsity profile, allowing for different nonzero weights. A group extension of the ℓ_0 problem asks for solutions of maximal group sparsity, where coefficients $\boldsymbol{s}_i$ are eliminated jointly. Extending our reweighted ℓ_1 ARD algorithm accordingly, we obtain a method where inner minimizations are second-order-cone programs. Once more, empirical performance improvements over standard MAP relaxations can be substantial [22], and these observations are backed up by theoretical results [35].



[FIG7] MEG source localization simulation example. A signal y is obtained from MEG sensors on the scalp surface, measuring small magnetic fields induced by cortical current flow. The design X can be constructed via Maxwell's equations and a structural MRI scan. Note that currents/sources u (in red) are typically confined to compact regions of the brain for a given experiment, they are sparse. We compare ARD source localization with ℓ_1 MAP reconstruction (1) and minimum variance adaptive beamforming [31].

DISCUSSION

Bayesian methods differ from MAP point estimation in that the unknown signal is averaged over the posterior distribution rather than fixed to a single best guess. Conceptual advantages of Bayesian averaging over MAP are typically offset by the higher running costs and less rigorous theoretical characterizations of most commonly used approximate inference methods. With novel variational Bayesian inference techniques reviewed here, the efficiency gap relative to convex MAP reconstruction can be narrowed considerably. Applied to sparse reconstruction, they constitute attractive alternatives to convex or nonconvex separable MAP estimation. At the cost of calling reweighted ℓ_1 problems a few times, they come with provable advantages in important applications, where restricted isometry properties typically do not hold. Moreover, they allow to quantify prediction confidences and uncertainty structure that can be used for advanced decision making problems beyond point reconstruction, such as automatic acquisition optimization by Bayesian experimental design. Their associated optimization problems are well characterized and can be solved efficiently even for very large models by way of scalable double-loop algorithms, reducing approximate Bayesian inference to separable point reconstruction and Gaussian graphical model computations coming together. Within the methodology reviewed here, Bayesian graphical models gain surprising new relevance to help to improve, as well as broaden, the scope of compressive sensing.

ACKNOWLEDGMENTS

Matthias W. Seeger thanks Hannes Nickisch and Florian Steinke for kindly providing figures. He also thanks Hannes Nickisch for discussions and comments. Matthias W. Seeger acknowledges funding by the Excellence Initiative of the German research foundation (DFG). David P. Wipf is supported by an NIH postdoctoral fellowship, award number F32NS061395 from the National Institute of Neurological Disorders and Stroke.

AUTHORS

Matthias W. Seeger (mseeger@mmci.uni-saarland.de) received the Ph.D. degree in informatics from the University of Edinburgh, United Kingdom, in 2003. He was a research associate at the University of California at Berkeley (2003–2005) and at the Max Planck Institute, Tübingen, Germany (2005–2008). He leads a research group at Saarland University and the Max Planck Institute, Saarbrücken, Germany. His research interests include probabilistic machine learning, Gaussian process methods, approximate Bayesian inference, applications to imaging and low level computer vision, and PAC-Bayesian analysis of nonparametric statistical methods.

David P. Wipf (dwipf@radiology.ucsf.edu) received the B.S. degree in electrical engineering from the University of Virginia and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California, San Diego. Currently, he is an NIH Postdoctoral Fellow in the Biomagnetic Imaging Lab, University of California, San Francisco. His research involves the development and analysis of Bayesian learning algorithms for functional brain imaging and sparse coding.

REFERENCES

[1] M. Wainwright, E. Simoncelli, and A. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Appl. Comput. Harmon. Anal.*, vol. 11, no. 1, pp. 89–123, 2001.

[2] A. Bruckstein, D. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, 2009.

[3] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[4] M. Seeger, "Bayesian inference and optimal design for the sparse linear model," *J. Mach. Learn. Res.*, vol. 9, pp. 759–813, Apr. 2008.

[5] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao, "Variational EM algorithms for non-Gaussian latent variable models," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 1059–1066.

[6] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1–2, pp. 1–305, 2008.

[7] A. Dalalyan and A. Tsybakov, "Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity," *Mach. Learn.*, vol. 72, no. 1–2, pp. 39–61, 2008.

[8] M. Seeger and H. Nickisch, "Compressed sensing and Bayesian experimental design," in *Proc. Int. Conf. Machine Learning 25*, A. McCallum, S. Roweis, and R. Silva, Eds. Omni Press, 2008, pp. 912–919.

[9] L. He and L. Carin, "Exploiting structure in wavelet-based Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 57, no. 9, pp. 3488–3497, 2009.

[10] M. I. Jordan, Ed., *Learning in Graphical Models*. Norwell, MA: Kluwer, 1997.

[11] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Comput.*, vol. 13, no. 11, pp. 2517–2532, 2001.

[12] D. Malioutov, J. Johnson, and A. Willsky, "Walk-sums and belief propagation in Gaussian graphical models," *J. Mach. Learn. Res.*, vol. 7, pp. 2031–2064, Oct. 2006.

[13] M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf, "Bayesian experimental design of magnetic resonance imaging sequences," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. New York: Curan Associates, 2009, pp. 1441–1448.

[14] M. Seeger and H. Nickisch, "Large scale variational inference and experimental design for sparse generalized linear models," *MPI Biological Cybernetics*, Tübingen, Germany, Tech. Rep. 175, Sept. 2008.

[15] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. New York: Curan Associates, 2008, pp. 1625–1632.

[16] H. Nickisch and M. Seeger, "Convex variational Bayesian inference for large scale generalized linear models," in *Proc. Int. Conf. Machine Learning 26*, L. Bottou and M. Littman, Eds. Omni Press, 2009, pp. 761–768.

[17] D. Malioutov, J. Johnson, M. Choi, and A. Willsky, "Low-rank variance estimation in GMRF models: Single and multiscale approaches," *IEEE Trans. Signal Processing*, vol. 56, no. 10, pp. 4621–4634, 2008.

[18] M. Schneider and A. Willsky, "Krylov subspace estimation," *SIAM J. Scientific Comput.*, vol. 22, no. 5, pp. 1840–1864, 2001.

[19] A. Willsky, "Multiresolution Markov models for signal and image processing," *Proc. IEEE*, vol. 8, no. 90, pp. 1396–1458, 2002.

[20] M. Wainwright, E. Sudderth, and A. Willsky, "Tree-based modeling and estimation of Gaussian processes on graphs with cycles," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001, pp. 661–667.

[21] M. Seeger, "Gaussian covariance and scalable variational inference," in *Proc. Int. Conf. Machine Learning 27*, J. Fürnkranz and T. Joachims, Eds. Omni Press, 2010.

[22] D. Wipf and S. Nagarajan, "Iterative reweighted $\ell 1$ and $\ell 2$ methods for finding sparse solutions," *IEEE J. Select. Topics in Signal Processing,* vol. 4, no. 2, pp. 317–329, 2010.

[23] D. Wipf and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *IEEE Trans. Inform. Theory,* submitted for publication.

[24] G. Wright, "Magnetic resonance imaging: From basic physics to imaging principles," *IEEE Signal Processing Mag.*, vol. 14, no. 1, pp. 56–66, 1997.

[25] M. Lustig, D. Donoho, and J. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, 2007.

[26] M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf, "Optimization of k-space trajectories for compressed sensing by Bayesian experimental design," *Magn. Reson. Med.*, vol. 63, no. 1, pp. 116–126, 2010.

[27] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," *Statist. Sci.*, vol. 10, no. 3, pp. 273–304, 1995.

[28] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Selective sampling for sparse signal recovery," in *Proc. Workshop on Artificial Intelligence and Statistics 12*, D. van Dyk and M. Welling, Eds. 2008, pp. 216–223.

[29] M. Seeger, "Speeding up magnetic resonance image acquisition by Bayesian multi-slice adaptive compressed sensing," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds. New York: Curan Associates, 2009, pp. 1633–1641.

[30] D. Manolakis, V. Ingle, and S. Kogon, *Statistical and Adaptive Signal Processing*. New York: McGraw-Hill, 2000.

[31] D. Wipf, J. Owen, H. Attias, K. Sekihara, and S. Nagarajan, "Robust Bayesian estimation of the location, orientation, and timecourse of multiple correlated neural sources using MEG," *NeuroImage*, vol. 49, no. 1, pp. 641–655, 2010.

[32] D. Wipf and S. Nagarajan, "Beamforming using the relevance vector machine," in *Proc. Int. Conf. Machine Learning 24,* Z. Ghahramani, Ed. Omni Press, 2007, pp. 1023–1030.

[33] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Statist. Soc. B*, vol. 68, no. 1, pp. 49–67, 2006.

[34] J. Tropp, "Algorithms for simultaneous sparse approximation—Part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, 2006.

[35] D. Wipf and B. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Signal Processing*, vol. 55, no. 7, pp. 3704–3716, 2007.