Sparse linear models: Variational approximate inference and Bayesian experimental design

Matthias W Seeger

Saarland University and Max Planck Institute for Informatics, Campus E1.4, 66123
Saarbrücken, Germany

Abstract. A wide range of problems such as signal reconstruction, denoising, source separation, feature selection, and graphical model search are addressed today by posterior maximization for linear models with sparsity-favouring prior distributions. The Bayesian posterior contains useful information far beyond its mode, which can be used to drive methods for sampling optimization (active learning), feature relevance ranking, or hyperparameter estimation, if only this representation of uncertainty can be approximated in a tractable manner. In this paper, we review recent results for variational sparse inference, and show that they share underlying computational primitives. We discuss how sampling optimization can be implemented as sequential Bayesian experimental design. While there has been tremendous recent activity to develop sparse estimation, little attendance has been given to sparse approximate inference. In this paper, we argue that many problems in practice, such as compressive sensing for real-world image reconstruction, are served much better by proper uncertainty approximations than by ever more aggressive sparse estimation algorithms. Moreover, since some variational inference methods have been given strong convex optimization characterizations recently, theoretical analysis may become possible, promising new insights into nonlinear experimental design.

1. Introduction

Most real-world signal classes of interest are highly structured. While precise high level statistical modelling of such structure is intractable at present, it imprints robust traces onto low level representations, which are in the feasible range of today's statistical methodology. For example, an image measured by a digital camera or a magnetic resonance scanner exhibits coherent edges, textures, and smooth areas, revealed by applying derivative or wavelet filters to the image bitmap: the filter responses show remarkably robust statistical behaviour far from Gaussianity (which would be obtained by filtering random noise) [1]. This low level statistical understanding, known as *sparsity* of image statistics, is used widely today in order to compress image bitmaps and video frame sequences. Similarly robust deviations from high-entropy Gaussianity occur for speech or music audio waveforms, leading to efficient compression schemes which can be decoded in real time. The statistical mechanisms underlying such low level applications are sparse linear models. The predominance of this technology for direct applications, such as coding and decoding (estimation), highlights the relevance of formulating and developing Bayesian frameworks, with which higher-order problems become accessible. In this paper, we will outline such a program, reviewing recent progress, as well as pointing out directions for research.

A running example in this paper will be sampling design optimization for magnetic resonance imaging (MRI) [2, 3], which has recently been addressed in [4]. In MRI, Fourier coefficients of selected image slices are acquired in a sequential fashion: if images of desired quality can be reconstructed from fewer acquisitions, we achieve a reduction in scan time, arguably today's most relevant limitation of MRI technology. Iterative nonlinear *compressive sensing* methods can be used to robustly undercut the classical Nyquist-Shannon sampling limit [5, 6, 7, 8]. The statistical idea behind these techniques is to import knowledge about "sparse" low level image statistics by way of a prior distribution, with which redundancies of dense sampling applied to images are exposed. Within this line of work, two major questions have to be faced: how to reconstruct given a fixed sampling design, and how to choose the design in the first place? The first question is addressed by *sparse estimation*. Given fixed design and data, a single most sensible image is produced by way of point estimation, minimizing a criterion composed of a data fit and a sparsity prior (or regularization) term. A wide array of convex and non-convex sparse estimation algorithms have been proposed. The second question, sampling design optimization, is different and has seen little attention so far, while there is evidence that the choice of design can be more relevant for successful real-world image reconstruction than the estimation algorithm or sparsity penalization used [9].

The nonlinear design optimization problem for realistic images has not been optimally solved, nor has a fully satisfying characterization been provided. Recent compressive sensing theory [5, 6] applies to unstructured, highly exactly sparse signals, but entirely fails to support design choices for natural or real medical images [9, 10, 4]. Short of optimal answers and conclusive theory, it is sensible to address the problem in an adaptive, data-driven manner, using concepts from Bayesian machine learning [4]. In this article, design optimization is phrased as search over realizable sampling patterns, optimizing a statistical criterion based on prior knowledge and real-world training data, with the aim of tractably detecting a pattern that generalizes well to future reconstruction problems. How could such a criterion look like? Which information from the data should it be based on? Which tools are available today in order to tractably evaluate it along the search? In this paper, we argue that these questions *can* successfully and tractably be addressed by adopting an approximate Bayesian viewpoint, extending the presently prevailing sparse estimation approaches by representations of *uncertainty*. In order to improve a given sampling design, a natural idea is to quantify the present shortcomings on real training data in a way that allows scoring additional candidate measurements as to how much novel information they are able to provide. This is not an image reconstruction problem per se, and it is hard to see how sparse point estimation techniques could be used to address it. Rather, it is an instance of nonlinear *Bayesian experimental design*, if the uncertainty representation in this picture is provided for by the Bayesian posterior, a distribution over all possible image reconstructions. Novel variational inference approximations can be used in order to implement this idea in MRI practice. Sparse *estimation*, reporting the mode of the Bayesian posterior, is not well suited to address higher-order design optimization. The latter is successfully driven by sparse *inference*, quantifying shape and covariance of the posterior beyond its mode.

At present, approximate Bayesian inference is hardly ever applied to low level imaging problems[1], for which much more efficient and well-characterized point estimation technology is preferred. Variational inference technology for continuous-variable image models is developed mainly in machine learning or Bayesian statistics. Since issues such as scalability, numerical robustness, or reductions to standard optimization primitives are not paid much attention there, previous methods run orders of magnitude slower than reconstruction techniques on the same model, so that sampling optimization for full high-resolution images cannot sensibly be addressed with them. A second goal of this article is to raise awareness of these issues. We show that at least some variational Bayesian approximations can be solved by scalable, well-characterized algorithms, which reduce to underlying standard image computations. As reviewed

[1] Bayesian *sparse estimation* techniques *are* applied to such problems, for example sparse Bayesian learning [11]. The crucial differences to Bayesian *sparse inference* will be highlighted below.

here, the major variational inference relaxations known today share similarities, and scalable reformulations should be attempted for all of them. Approximate Bayesian inference applied to low level real-world problems may come with new concepts and help to overcome obstacles in novel ways. Yet being based on the same model, prior assumptions and basic primitives, it should not have to come with entirely new computational methodology. If convex optimization and numerical mathematics technology is highly successful for imaging, in practice as well as in theoretical understanding, higher order Bayesian methods would ideally make use of these advances directly, instead of going different uncharted ways altogether.

The structure of the paper is as follows. The sparse linear model is introduced in Section 2, along with Bayesian nonlinear experimental design, motivated by MRI sampling design optimization. Current variational approximations to sparse Bayesian inference are reviewed in Section 3, along with convexity characterizations and scalable algorithms. This paper draws on material from previous publications, notably [12, 9, 4, 13], and on the report [14].

2. Sparse linear model. Experimental design

Let us formulate the setup for the MRI sampling optimization problem. Denote the desired MR image by $\boldsymbol{u} \in \mathbb{R}^n$, where n is the number of pixels. Under ideal conditions, the raw data $\boldsymbol{y} \in \mathbb{R}^m$ from an NMR scanner consists of Fourier coefficients of $\boldsymbol{u}$ at specific k-space spatial frequencies, so that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \boldsymbol{\varepsilon}$, $\quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Rows of $\boldsymbol{X}$ are Fourier filters (k-space is synonimous with 2D Fourier space)[2]. Elementary units of the design (blocks of rows of $\boldsymbol{X}$) are called *phase encodes*. For example, in Cartesian MRI, a phase encode is a complete column in k-space. A Nyquist-dense design contains encodes to cover k-space completely, while for an undersampled design $\boldsymbol{X}$, certain encodes are not acquired ($m < n$). For undersampled reconstruction, a prior $P(\boldsymbol{u})$ is chosen which represents low level statistics of (MR) images, distinctly super-Gaussian "sparse" distributions. The posterior has the form

$$P(\boldsymbol{u}|\boldsymbol{y}) \propto N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) \prod_{j=1}^{q} e^{-\tau_j |s_j/\sigma|}, \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}. \tag{1}$$

The likelihood $P(\boldsymbol{y}|\boldsymbol{u}) = N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I})$ is Gaussian, while the prior $P(\boldsymbol{u})$ is the normalized product of q Laplacians on linear projections s_j of $\boldsymbol{u}$, among them the image gradient and wavelet coefficients (see [9] for details), so that typically $q > n$. The Laplace distribution encourages sparsity of $\boldsymbol{s}$ [12]. This posterior is log-concave: it has a single mode, and all level sets are convex. We refer to this setup as *sparse linear model* (SLM). Note that we omit (parts of) $\boldsymbol{X}$ from distribution notation. These are control variables (covariates), and distributions are conditioned on them implicitly[3].

This Bayesian setup can be used for *sparse estimation* or *sparse inference* respectively. Sparse estimation concerns the reconstruction of a single image, which is often done by maximizing the posterior, or *MAP estimation*: $\hat{\boldsymbol{u}} = \operatorname{argmin}_{\boldsymbol{u}} -\log P(\boldsymbol{y}|\boldsymbol{u}) - \log P(\boldsymbol{u})$. For our setup here, this is a convex quadratic program, a special case of which ($\boldsymbol{B} = \boldsymbol{I}$) is known as the Lasso [15]. *Sparse inference* goes beyond finding the posterior mode, trying to approximate posterior moments such as mean $\mathrm{E}_P[\boldsymbol{u}|\boldsymbol{y}]$ or covariance $\mathrm{Cov}_P[\boldsymbol{u}|\boldsymbol{y}]$, or its log-partition function $\log P(\boldsymbol{y})$. Reviewing the numerous good reasons for doing so is not in the scope of this paper (see for example [16]): we will concentrate solely on design optimization. We will see below that Bayesian

[2] In practice, both $\boldsymbol{u}$ and $\boldsymbol{y}$ are complex-valued, which was neglected (for $\boldsymbol{u}$) in [4], but is accounted for properly [14] by doubling the number of real-valued variables (using the $\mathbb{C} \to \mathbb{R}^2$ embedding), and placing sparsity potentials on $|s_j|$ instead of s_j (potentials are even functions). This issue, which does not add implementational complexity, is ignored in the remainder of this paper, while it is crucial for making the MRI application work. More generally, norm potentials can be placed on $\|\boldsymbol{s}_j\|$ [14], using minor modifications.

[3] It is useful to address uncertainties in $\boldsymbol{X}$ itself, but this is not done here.

experimental design in our setting here is mainly driven by the posterior covariance, largely unrelated to the mode $\hat{\boldsymbol{u}}$. It is widely accepted today that the computational simplicity of finding the latter does not transfer to Bayesian inference, which remains intractable even for log-concave posteriors. Approximate sparse inference can be done by Markov chain Monte Carlo [17] or by variational approximations, both originally proposed in statistical physics. While most Bayesian statisticians prefer to use MCMC, little is known about how to obtain rigorous finite-time, limited-resources performance guarantees. In this paper, we concentrate on inference approximations which can be relaxed to image reconstruction technology in place already (such as convex optimization), thus on variational approximations.

In the remainder of this section, we briefly review Bayesian sequential experimental design and motivate how it can be used to address MRI sampling optimization. More general reviews are found in [18, 19, 20]. Recalling the basic idea from Section 1, we have to score extension candidates $\boldsymbol{X}_* \in \mathbb{R}^{d,n}$ (phase encodes) with respect to the decision of appending them to the present design $\boldsymbol{X}$. A generally useful measure is the *information gain*

$$\Delta(\boldsymbol{X}_*) := \mathrm{H}[P(\boldsymbol{u}|\boldsymbol{y})] - \mathrm{E}_{P(\boldsymbol{y}_*|\boldsymbol{y})}\left[\mathrm{H}[P(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{y}_*)]\right], \tag{2}$$

where $P(\boldsymbol{u}|\boldsymbol{y})$ is the Bayesian posterior before, $P(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{y}_*)$ after including $(\boldsymbol{X}_*,\boldsymbol{y}_*)$, $P(\boldsymbol{y}_*|\boldsymbol{y}) = \mathrm{E}_{P(\boldsymbol{u}|\boldsymbol{y})}[P(\boldsymbol{y}_*|\boldsymbol{u})]$ is the predictive posterior, and $\mathrm{H}[P(\boldsymbol{u}|\boldsymbol{y})] := \mathrm{E}_{P(\boldsymbol{u}|\boldsymbol{y})}[-\log P(\boldsymbol{u}|\boldsymbol{y})]$ is the Shannon (differential) entropy [21]. The information gain measures the decrease in uncertainty about $\boldsymbol{u}$, averaged over the "soft" prediction $P(\boldsymbol{y}_*|\boldsymbol{y})$. It is not the only criterion that can be used [20], especially if more specific utilities can be defined for the task at hand. What *is* characteristic of Bayesian experimental design, is that averages over $P(\boldsymbol{u}|\boldsymbol{y})$ and $P(\boldsymbol{y}_*|\boldsymbol{y})$ are taken: a principal and important difference to plugging in "best estimates", the preferred option with purely estimation-based techniques. The highest-scoring encode is appended to $\boldsymbol{X}$, and the process is iterated sequentially. When applying this procedure to MRI sampling optimization [4], many candidates $\boldsymbol{X}_*$ can be scored in each round without performing costly NMR measurements for them: an important advantage over "trial-and-error" approaches.

Given that we have to approximate sparse inference, it is important to understand *which* posterior moments are required in order to evaluate the information gain (2). In this paper, we concentrate on Gaussian posterior approximations $Q(\boldsymbol{u}|\boldsymbol{y}) \approx P(\boldsymbol{u}|\boldsymbol{y})$, fitted by means of a variational optimization problem. For a Gaussian, $\mathrm{H}[Q(\boldsymbol{u}|\boldsymbol{y})] = (1/2)\log|\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}]| + C$. Moreover, we approximate $P(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{y}_*)$ by the Gaussian $\propto N(\boldsymbol{y}_*|\boldsymbol{X}_*\boldsymbol{u},\sigma^2\boldsymbol{I})Q(\boldsymbol{u}|\boldsymbol{y})$, so to avoid having to re-run the variational optimization for each candidate $\boldsymbol{X}_*$. With these approximations,

$$\Delta(\boldsymbol{X}_*) \approx -\log|\boldsymbol{A}| + \log\left|\boldsymbol{A} + \boldsymbol{X}_*^T\boldsymbol{X}_*\right| = \log\left|\boldsymbol{I} + \boldsymbol{X}_*\boldsymbol{A}^{-1}\boldsymbol{X}_*^T\right|, \quad \mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}] = \sigma^2\boldsymbol{A}^{-1}. \tag{3}$$

The notation in terms of $\boldsymbol{A}$ will be convenient below. Therefore, for the methods of interest here, $\Delta(\boldsymbol{X}_*)$ most strongly depends on the posterior covariance $\mathrm{Cov}_P[\boldsymbol{u}|\boldsymbol{y}]$. More precisely, (3) is dominated by the largest eigenvectors and eigenvalues of $\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}]$, which fits into the intuitive picture of Section 1: we assess the directions of largest posterior spread of uncertainty, then align novel measurements with these directions of maximum uncertainty in what $\boldsymbol{u}$ should be. While we do not even know how to represent the exact SLM posterior, its leading covariance eigendirections constitute a definite approximation target in practice as well as for theoretical investigations.

3. Variational inference for sparse linear models

In this section, we review variational approximation techniques which have been applied to Bayesian inference for the sparse linear model. We will be interested in properties of the underlying variational optimization problems which allows for reductions to commonly used scientific computing techniques. We rely on [22, 12].

The posterior $P(\boldsymbol{u}|\boldsymbol{y})$ is intractable to integrate over for two reasons coming together. First, $\boldsymbol{u}$ is very high-dimensional, so that black-box techniques like quadrature are impossibly expensive. Second, $P(\boldsymbol{u}|\boldsymbol{y})$ is not Gaussian, due to the presence of non-Gaussian potentials $t_j(s_j) = e^{-\tau_j|s_j/\sigma|}$. Gaussian integrals are analytically tractable linear algebra expressions, which can often be approximated reliably even at large scales, by exploiting structure in the matrices $\boldsymbol{X}$, $\boldsymbol{B}$. Variational inference techniques relax the problem of integrating against $P(\boldsymbol{u}|\boldsymbol{y})$, by replacing the posterior (or relevant moments thereof) by approximations which allow for tractable computation. This goes hand in hand with introducing new variational parameters, and the variational problem is to optimize over these in order to fit the approximation to the posterior. Once this is done, the posterior approximation is used in place of the true posterior, when it comes to Bayesian queries such as design score computations (2). In this paper, we concentrate on *Gaussian* posterior approximations $Q(\boldsymbol{u}|\boldsymbol{y}) \approx P(\boldsymbol{u}|\boldsymbol{y})$, of the form

$$Q(\boldsymbol{u}|\boldsymbol{y}) = N(\boldsymbol{u}|\boldsymbol{u}_*, \sigma^2\boldsymbol{A}^{-1}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_{j=1}^{q} e^{-\sigma^{-2}(s_j^2/(2\gamma_j)-b_js_j)}, \quad \boldsymbol{A} := \boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{B}, \qquad (4)$$

where $\boldsymbol{\Gamma} := \operatorname{diag}\boldsymbol{\gamma}$. The family $\{Q(\boldsymbol{u}|\boldsymbol{y})\}$ is formally obtained from $P(\boldsymbol{u}|\boldsymbol{y})$ by replacing potentials $t_j(s_j)$ with Gaussian functions, introducing variational parameters $\boldsymbol{\gamma} = (\gamma_j)$, $\boldsymbol{b} = (b_j)$. Most of the methods discussed below do not impose the structure of (4), but rather start with a general idea of how to approximate $P(\boldsymbol{u}|\boldsymbol{y})$. Given that, the form (4) is *implied*.

What is the role of the variational parameters? For sparse linear models, the $\boldsymbol{b}$ parameters seem much less relevant. In some techniques, they are fixed to zero or other constant values. For other models not discussed here, $\boldsymbol{b}$ plays a more significant role (they allow to shift the mean without affecting the covariance). The variance parameters $\boldsymbol{\gamma}$ are instrumental. If $\boldsymbol{\gamma} \succeq \boldsymbol{0}$, it is easy to show that[4] $\mathrm{Var}_Q[s_j|\boldsymbol{y}] \le \sigma^2\gamma_j$ [14], so that the approximate posterior variance along s_j is bounded in terms of γ_j. Sparse linear models embody *selective shrinkage*: most $|s_j|$ are shrunk strongly towards zero, while some $|s_j|$ are hardly penalized at all. This property captures low level statistical properties of real-world signals discussed above[5]. Selective shrinkage is controlled by the γ_j parameters, which will be small for coefficients that are shrunken under the posterior, but large for such that are not strongly penalized. Readers familiar with image modelling may be sceptical about the proposal to approximate sparse image model posteriors by Gaussians. Is it not generally accepted that images have strong non-Gaussian statistics, for example along edges? Is it not a main argument of this paper that non-Gaussian linear models are to be used? This “dilemma” is solved by the presence of variational parameters. While $Q(\boldsymbol{u}|\boldsymbol{y})$ is a Gaussian, it is heavily parameterized (typically, q is larger than n) in a very non-stationary manner. While an “average” member of $\{Q(\boldsymbol{u}|\boldsymbol{y})\}$ may not represent image statistics, the closest fit to $P(\boldsymbol{u}|\boldsymbol{y})$ from this large family does so indeed[6].

How will we approximate the posterior covariance matrix $\mathrm{Cov}_P[\boldsymbol{u}|\boldsymbol{y}]$? In contrast to discrete variable variational relaxations, where only very simple distributions can be represented exactly, the Gaussian family $\{Q(\boldsymbol{u}|\boldsymbol{y})\}$ is rich enough to represent a wide range of *global* covariances, to the extent that for a log-concave (uni-modal) posterior $P(\boldsymbol{u}|\boldsymbol{y})$, the covariance of a close global fit $Q(\boldsymbol{u}|\boldsymbol{y}) \approx P(\boldsymbol{u}|\boldsymbol{y})$ can be a useful proxy for the leading directions of posterior covariance (see also end of Section 3.4).

[4] Strictly speaking, the proof assumes that $\boldsymbol{\gamma} \succ \boldsymbol{0}$. By continuity, it holds also if coefficients of $\boldsymbol{\gamma}$ become zero, as long as $\boldsymbol{A}$ stays positive definite.

[5] These properties are often referred to as *sparsity* (hence sparse linear model), but *super-Gaussianity* (introduced below) would be a better term in the context of natural images, which are *not* piecewise constant (which sparsity would imply), but filter coefficients exhibit a super-Gaussian (power law) decay.

[6] If $\boldsymbol{\gamma}$ for this closest fit is plotted accordingly for large enough m, edges of the underlying image are typically revealed.

3.1. Super-Gaussianity. Scale mixtures. Direct site bounding

How to replace single potentials $t_j(s_j)$ by Gaussian functions in a principled manner? Two general ideas are reviewed in [22]: writing $t_j(s_j)$ as a *scale mixture*, and lower-bounding $t_j(s_j)$ by a Gaussian function. In the former case, $t_j(s_j) = \int_{>0} N(s_j|0, \sigma^2\gamma_j) f_j(\gamma_j)\, d\gamma_j$. In the latter case, $t_j(s_j) = \max_{\gamma_j>0} e^{-(s_j/\sigma)^2/(2\gamma_j) - h_j(\gamma_j)/2}$. Both ideas are restricted to potentials $t_j(s_j)$ for which such representations exist [22]. Potentials for which the latter max-representation exists, are called *super-Gaussian*: they can be lower bounded by Gaussian functions of any width. This definition is equivalent to $t_j(s_j)$ being even, positive, and $s_j^2 \mapsto \log t_j(s_j)$ being convex and decreasing. Namely, if $x_j := (s_j/\sigma)^2$, $\lambda_j := -1/(2\gamma_j)$, then $\log t_j(s_j) = \max_{\lambda_j<0} x_j\lambda_j - h_j(-1/(2\lambda_j))/2$, convex as maximum over affine functions (this is a Legendre/Fenchel duality, see [23]). Many potential functions used in statistics are super-Gaussian. Remarkably, it is shown in [22] that all scale mixture potentials are super-Gaussian: the latter concept is more general. For Laplacian potentials (which are scale mixtures), $h_j(\gamma_j) = \tau_j^2\gamma_j$. Further examples and closure properties can be found in [22, 14]. In the context of SLMs, super-Gaussianity is precisely the property associated with sparsity-favouring potentials: for a Gaussian $t_j(s_j)$, $s_j^2 \mapsto \log t_j(s_j)$ is affine, so that convexity of this function translates into heavier tails and stronger concentration of mass close to zero. The concept can be extended to non-even potentials, whenever $t_j(s_j)e^{\kappa_j s_j}$ is even for some κ_j. For example, Bernoulli (or logistic) potentials used as binary classification likelihoods are super-Gaussian, which was noted and exploited in [24].

An early Bayesian treatment of SLMs was sparse Bayesian learning (SBL) [11], motivated in terms of scale mixture decompositions for Student's t potentials, but derived in a somewhat heuristic manner. The maximum a posteriori arguments used to motivate SBL are problematic, because by changing the parameterization (say, γ_j versus $\pi_j = 1/\gamma_j$), we arrive at different algorithms, which may address different problems (sparse estimation, sparse inference), and whose convergence properties and speed can be very different. SBL, which in the form applied in [11] is a sparse *estimation* technique, has found widespread use, in the wake of which the concepts of different variational relaxation principles, sparsity priors, and update algorithms have been mixed in a confusing manner. Confusions of this kind can be avoided by focussing on approximation *principles* (independent of *algorithms* for solving them) that are invariant to parameterizations, which is what we restrict ourselves to in this paper. Among sparse estimation techniques, SBL (also known as ARD) works remarkably well, often improving on MAP estimation significantly [25].

For super-Gaussian potentials $t_j(s_j)$, *direct site bounding*[7] (DSB) is obtained by plugging site bounds into the log partition function:

$$\log P(\boldsymbol{y}) \geq \max_{\boldsymbol{\gamma}\succ\mathbf{0}} -\phi_{\text{DSB}}(\boldsymbol{\gamma})/2, \quad \phi_{\text{DSB}}(\boldsymbol{\gamma}) := -2\log Z_Q + h(\boldsymbol{\gamma}),$$
$$Z_Q := \int P(\boldsymbol{y}|\boldsymbol{u}) e^{-\frac{1}{2}\sigma^{-2}\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}}\, d\boldsymbol{u}, \quad \boldsymbol{\Gamma} := \operatorname{diag}\boldsymbol{\gamma},\ h(\boldsymbol{\gamma}) := \sum_j h_j(\gamma_j). \tag{5}$$

The variational problem is $\min_{\boldsymbol{\gamma}\succ\mathbf{0}} \phi_{\text{DSB}}(\boldsymbol{\gamma})$, motivated by tightening the lower bound to $\log P(\boldsymbol{y})$. Why should this lead to a closer fit of $Q(\boldsymbol{u}|\boldsymbol{y})$ to $P(\boldsymbol{u}|\boldsymbol{y})$? The log partition function $\log P(\boldsymbol{y})$ is the general approximation target of variational techniques. It is the moment generating function of $P(\boldsymbol{u}|\boldsymbol{y})$, while the lower bound plays the same role for $Q(\boldsymbol{u}|\boldsymbol{y})$. A more direct interpretation in terms of relative entropy divergence is given in Section 3.2. However, at present there is no finite-size theory we are aware of that firmly links improvements in lower bound tightness to approximation quality of posterior moments. Note that Z_Q is the normalization constant of $Q(\boldsymbol{u}|\boldsymbol{y})$, whose form does not depend on specifics of the potentials $t_j(s_j)$.

[7] As opposed to SBL, DSB is not a commonly used terminology, but was chosen for convenience here.

As a Gaussian integral, $\phi_{\mathrm{DSB}}(\boldsymbol{\gamma})$ can be evaluated tractably. A simple optimization strategy is coordinate descent, updating each γ_j in turn while keeping all others fixed. For Laplacian potentials, this algorithm is due to [26], see [22, 12] for more general discussions. The update for γ_j is a simple function of the posterior marginal $Q(s_j|\boldsymbol{y})$. As shown in [22], this algorithm can be implemented without knowing $h_j(\gamma_j)$ explicitly. However, in order to update all parameters, posterior means *and* variances are required for all s_j. In cases where the precision matrix $\sigma^{-2}\boldsymbol{A}$ of $Q(s_j|\boldsymbol{y})$ exhibits a sparse graphical model structure, such *single site updating* algorithms can be implemented by message passing. Once a potential has been updated, the information is propagated to the next site to be visited. In large scale situations of interest here, where $Q(s_j|\boldsymbol{y})$ does not have useful graphical model structure, a $n \times n$ linear system has to be solved from scratch in order to access any single marginal $Q(s_j|\boldsymbol{y})$: message passing is not attractive. We will come back to this point in Section 3.4.

3.2. Variational mean field approximations

A variational mean field (VMF)[8] characterization of Bayesian inference is given by

$$\log P(\boldsymbol{y}) = \max_{Q(\boldsymbol{u}|\boldsymbol{y})} \mathrm{E}_Q\left[\log P(\boldsymbol{y},\boldsymbol{u}) - \log Q(\boldsymbol{u}|\boldsymbol{y})\right], \tag{6}$$

where the maximum is taken over all distributions [27]. The maximizer is given by $Q(\boldsymbol{u}|\boldsymbol{y}) = P(\boldsymbol{u}|\boldsymbol{y})$, the true posterior. The slack in the lower bound for any fixed $Q(\boldsymbol{u}|\boldsymbol{y})$ is the relative entropy $\mathrm{D}[Q(\boldsymbol{u}|\boldsymbol{y}) \,\|\, P(\boldsymbol{u}|\boldsymbol{y})]$. VMF relaxations are obtained by restricting the maximum to a subset of distributions, for which the lower bound to $\log P(\boldsymbol{y})$ can be evaluated tractably. In other words, the closest member $Q(\boldsymbol{u}|\boldsymbol{y})$ from this subset in terms of $\mathrm{D}[\cdot \,\|\, P(\boldsymbol{u}|\boldsymbol{y})]$ divergence is pursued[9]. In this section, we will discuss two relaxations, which have been used for SLMs.

First, we may simply restrict ourselves to unconstrained *Gaussian* distributions $Q(\boldsymbol{u}|\boldsymbol{y})$, which allows for tractable lower bound maximization (this involves computing Gaussian expectations over $\log t_j(s_j)$). Interestingly, stationary points of this optimization have the form prescribed by (4) (see Appendix), so that $Q(\boldsymbol{u}|\boldsymbol{y})$ can be restricted to lie in this family without loss of generality. The lower bound becomes

$$\begin{aligned} &\log P(\boldsymbol{y}) \ge \max_{\boldsymbol{\gamma},\boldsymbol{b}} -\phi_{\mathrm{MF}}(\boldsymbol{\gamma},\boldsymbol{b})/2, \quad \phi_{\mathrm{MF}}(\boldsymbol{\gamma},\boldsymbol{b}) := -2\log Z_Q + \sum_j h_j^{\mathrm{MF}}(\gamma_j, b_j, Q(s_j|\boldsymbol{y})), \\ &h_j^{\mathrm{MF}}(\gamma_j, b_j, Q(s_j|\boldsymbol{y})) := -2\mathrm{E}_Q[\log t_j(s_j) + (s_j/\sigma)^2/(2\gamma_j) - b_j s_j/\sigma^2], \end{aligned} \tag{7}$$

where Z_Q is the normalization constant of $Q(\boldsymbol{u}|\boldsymbol{y})$. This criterion has a more complicated structure than (5), because h_j^{MF} depends on $Q(s_j|\boldsymbol{y})$. More can be said if $\boldsymbol{b}$ is fixed to zero and all $t_j(s_j)$ are super-Gaussian: $h_j^{\mathrm{MF}} = -2\mathrm{E}_Q[\log t_j(s_j) + (s_j/\sigma)^2/(2\gamma_j)]$. Since $\log t_j(s_j) + (s_j/\sigma)^2/(2\gamma_j) \ge -h_j(\gamma_j)/2$ for all s_j, $\gamma_j > 0$, we have that $h_j^{\mathrm{MF}}(\gamma_j, Q(s_j|\boldsymbol{y})) \le h_j(\gamma_j)$ for $\gamma_j > 0$. Therefore, this variant of VMF uses a potentially tighter bound on $\log P(\boldsymbol{y})$ than DSB does. Moreover, since $\log P(\boldsymbol{y}) = -2\phi_{\mathrm{MF}}(\boldsymbol{\gamma}) + \mathrm{D}[Q(\boldsymbol{u}|\boldsymbol{y}) \,\|\, P(\boldsymbol{u}|\boldsymbol{y})]$, this means that DSB minimizes an upper bound to the VMF relative entropy $\mathrm{D}[Q(\boldsymbol{u}|\boldsymbol{y}) \,\|\, P(\boldsymbol{u}|\boldsymbol{y})]$, which can have a simpler structure than the relative entropy itself (see Section 3.4).

[8] A historically more correct terminology would be variational *structured* mean field. In this paper, VMF approximations are those that make use of (6) with restricted families for $Q(\boldsymbol{u}|\boldsymbol{y})$. In general, the term VMF seems reserved for *factorization* restrictions only. In recent papers, structured VMF has been termed "variational Bayes", ignorant of the fact that methods like DSB or EP are variational Bayesian approximations just as well, but are based on different ideas.

[9] In general, VMF approximations do not constitute convex optimization problems. While the bound can be optimized locally, finding a global optimum may still be a hard problem.

A different variant of VMF is discussed in [22], for cases in which the $t_j(s_j)$ have scale mixture decompositions. Once $\boldsymbol{\gamma}$ are introduced alongside $\boldsymbol{u}$, the VMF characterization (6) can be written as maximum over $Q(\boldsymbol{u}, \boldsymbol{\gamma}|\boldsymbol{y})$ just as well, where $P(\boldsymbol{y}, \boldsymbol{u})$ is replaced by $P(\boldsymbol{y}|\boldsymbol{u})N(\boldsymbol{s}|\boldsymbol{0}, \sigma^2\boldsymbol{\Gamma})(f_j(\gamma_j))$ (recall the scale mixture decomposition of $t_j(s_j)$ from Section 3.1). In this case, tractability is obtained by making a *factorization* assumption: $Q(\boldsymbol{u}, \boldsymbol{\gamma}|\boldsymbol{y}) = Q(\boldsymbol{u}|\boldsymbol{y})Q(\boldsymbol{\gamma}|\boldsymbol{y})$. Once more, this free-form assumption implies further simplifications: $Q(\boldsymbol{u}|\boldsymbol{y})$ can be restricted to lie in the Gaussian family (4), with $\boldsymbol{b} = \boldsymbol{0}$. Moreover, it is shown in [22] that single site updating for this VMF variant is precisely equivalent to coordinate descent for DSB. Therefore, this second VMF variant is nothing new, but simply a different way of motivating the DSB variational problem.

3.3. Expectation propagation

The expectation propagation (EP) algorithm [28, 29] is based on the idea of moment matching. It builds on the assumed density filtering (ADF) principle [30]. Suppose a posterior of the form (1) is to be approximated by a Gaussian $Q(\boldsymbol{u}|\boldsymbol{y})$, where (different from the SLM discussed above) the Gaussian factor in $P(\boldsymbol{u}|\boldsymbol{y})$ is normalizable. In exact Bayesian filtering, we start with a Gaussian $Q_0(\boldsymbol{u})$ proportional to this factor, then include one potential $t_j(s_j)$ after the other, stepping from $Q_0(\boldsymbol{u})$ to $\propto Q_0(\boldsymbol{u})t_1(s_1)$, and so on. Even after the first inclusion, the filtering distribution is not Gaussian anymore, and computations become intractable rapidly. In ADF, after each Bayesian inclusion, we project the intermediate distribution back onto a Gaussian, preserving mean and covariance. This is equivalent to finding the solution $Q_1(\boldsymbol{u})$ of $\min_{Q \text{ Gaussian}} \mathrm{D}[\hat{P}_1(\boldsymbol{u}) \,\|\, Q(\boldsymbol{u})]$, where $\hat{P}_1(\boldsymbol{u}) \propto Q_0(\boldsymbol{u})t_1(s_1)$ (see Appendix). Moreover, to step from Q_0 to Q_1, we multiply the present state $Q_0(\boldsymbol{u})$ by $\tilde{t}_1(\boldsymbol{u}) \propto Q_1(\boldsymbol{u})/Q_0(\boldsymbol{u})$. In fact, since t_1 depends on s_1 only, we may choose $\tilde{t}_1(\boldsymbol{u}) = \tilde{t}_1(s_1) := e^{\sigma^{-2}(b_1 s_1 - s_1^2/(2\gamma_1))}$, where b_1, γ_1 can be computed from $Q_0(s_1)$ and $t_1(s_1)$ only (see Appendix). The ADF update is a local computation for models with local potentials $t_j(s_j)$. Moreover, the Gaussian approximation $Q(\boldsymbol{u}|\boldsymbol{y}) = Q_q(\boldsymbol{u})$ lies within the family (4).

Once all potentials $t_j(s_j)$ have been included, ADF has to be terminated. EP extends ADF by the capability of re-visiting previous projections, in order to obtain a full-fledged variational inference technique. If ADF updates were continued after the first round, we would overcount the influence of non-Gaussian potentials. An ADF update at $t_j(s_j)$ is done for a state $Q(\boldsymbol{u}|\boldsymbol{y})$ where $b_j = 1/\gamma_j = 0$. In order to re-create this situation, an EP update starts with dividing out the Gaussian term $\tilde{t}_j(s_j) := e^{\sigma^{-2}(b_j s_j - s_j^2/(2\gamma_j))}$, followed by an ADF update. In other words, the *cavity distribution* $Q_{\backslash j}(s_j|\boldsymbol{y}) \propto Q(s_j|\boldsymbol{y})\tilde{t}_j(s_j)^{-1}$ is used as basis for computing $\hat{P}_j(s_j)$, in place of the marginal $Q(s_j|\boldsymbol{y})$. The EP algorithm is of the single site updating type (see Section 3.1), running EP updates in some ordering until convergence. At a stationary point, *moment consistency* is attained: all $\hat{P}_j(\boldsymbol{u})$ share their Gaussian moments with $Q(\boldsymbol{u}|\boldsymbol{y})$. This condition should be compared with the first VMF variant in Section 3.2, where $\mathrm{D}[Q(\boldsymbol{u}|\boldsymbol{y}) \,\|\, P(\boldsymbol{u}|\boldsymbol{y})]$ is minimized over Gaussians. Following the information-theoretic interpretation of relative entropy $\mathrm{D}[P_a \,\|\, P_b]$ (coding loss if P_b is used instead of the true P_a), the ordering of arguments is the wrong way around. The minimization of $\mathrm{D}[P(\boldsymbol{u}|\boldsymbol{y}) \,\|\, Q(\boldsymbol{u}|\boldsymbol{y})]$ is of course intractable: it is equivalent to finding the *exact* posterior mean and covariance. EP can be seen as using the relative entropy with arguments the right way around, yet replacing the posterior $P(\boldsymbol{u}|\boldsymbol{y})$ by surrogates $\hat{P}_j(\boldsymbol{u})$ ("one step away from Gaussian") for which the divergence can be computed. With EP, γ_j can become negative in general [28]. However, if all $t_j(s_j)$ are log-concave, all γ_j remain positive throughout [12].

What is the variational problem underlying the EP algorithm? Let us introduce additional parameters C_j, defined by moment matching conditions of zero-th order: $\mathrm{E}_{Q_{\backslash j}}[t_j(s_j)] =$

$\mathrm{E}_{Q_{\setminus j}}[C_j\tilde{t}_j(s_j)]$. Replacing $t_j(s_j)$ by $C_j\tilde{t}_j(s_j)$, we obtain the approximation

$$\log P(\boldsymbol{y}) \approx -\phi_{\mathrm{EP}}(\boldsymbol{\gamma},\boldsymbol{b})/2, \quad \phi_{\mathrm{EP}}(\boldsymbol{\gamma},\boldsymbol{b}) := -2\log Z_Q - 2\sum_j \log C_j, \tag{8}$$

which is brought into the form previously used by defining $h_j^{\mathrm{EP}}(\gamma_j, b_j, Q(s_j|\boldsymbol{y})) := -2\log C_j$. Every fixed point of the EP algorithm is a stationary point of $\phi_{\mathrm{EP}}(\boldsymbol{\gamma},\boldsymbol{b})$ [12]. There are important differences to DSB and VMF. First, these stationary points are neither local maxima nor minima, but true saddle points (at least with respect to the unconstrained parameterization in terms of $(\boldsymbol{b},\boldsymbol{\gamma})$). The criterion $\phi_{\mathrm{EP}}(\boldsymbol{\gamma},\boldsymbol{b})$ has a more complicated structure than ϕ_{MF} or ϕ_{DSB}, and the EP algorithm does not always converge [28] (a convergence proof has not been given even for log-concave models). While there are provably convergent algorithms [31], they are significantly more expensive to run. Finally, EP seems more susceptible to numerical instability problems than the other techniques discussed here. For the SLM with (log-concave) Laplacian potentials, EP frequently fails on real-world image data [12]. Due to the underdetermined likelihood, cavity variances of $Q_{\setminus j}(s_j|\boldsymbol{y})$ tend to be huge, which leads to runaway numerical errors in the moment computations for $\hat{P}_j(s_j)$, even if these are executed with utmost care. In this case, the EP algorithm is highly sensitive to such errors. A remedy is to use *fractional EP* [32], in that $Q_{\setminus j}(s_j|\boldsymbol{y}) \propto Q(s_j|\boldsymbol{y})\tilde{t}(s_j)^{-\eta}$, $\hat{P}_j(s_j) \propto Q_{\setminus j}(s_j|\boldsymbol{y})t_j(s_j)^{\eta}$ for $\eta < 1$. While this constitutes a variational problem different[10] from standard EP, numerical problems essentially disappear even for $\eta = 0.9$ [12].

3.4. Convexity properties. Scalable double loop algorithms

We discussed a range of continuous-variable variational inference approximations (DSB, VMF, EP) above, along with single site updating algorithms. Implementations of these algorithms tend to be similar, and underlying variational criteria can be brought into similar form. This superficial similarity does not mean that the variational problems can be characterized or solved equally well in practice. We saw that EP is a saddle point rather than an optimization problem, which tends to be numerically more problematic than the others. In this section, we show that for log-concave posteriors (including the SLM introduced above), DSB is a convex optimization problem, and we provide an algorithm to solve it even for very large models. In comparison, VMF is a non-convex problem in general, and no scalable solvers have been proposed so far, let alone for EP.

All criteria share the part $-2\log Z_Q$. Suppose that $\boldsymbol{b}$ is fixed to zero or some other constant, so that $\boldsymbol{\gamma}$ are the sole variational parameters (natural for DSB, can be enforced for VMF, EP). $\boldsymbol{\gamma} \mapsto -2\log Z_Q$ is a convex function for $\boldsymbol{\gamma} \succ \boldsymbol{0}$ [14]. Namely, $\log P(\boldsymbol{y}|\boldsymbol{u}) + \sigma^{-2}(\boldsymbol{b}^T\boldsymbol{u} - \boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}/2)$ is jointly concave as function of $(\boldsymbol{u},\boldsymbol{\gamma})$, $\boldsymbol{\gamma} \succ \boldsymbol{0}$, so that $\boldsymbol{\gamma} \mapsto \log Z_Q$ is concave by Prékopa's marginalization theorem [33]. This holds only if $\boldsymbol{b}$ is constant: $\log Z_Q$ is not concave in $(\boldsymbol{b},\boldsymbol{\gamma})$ in general. It is shown in [14] that for super-Gaussian potentials, $h(\boldsymbol{\gamma})$ is convex iff all sites $t_j(s_j)$ are log-concave: the DSB *variational inference* problem is convex if and only if *MAP estimation* is convex for the same model. An equivalent characterization does not hold for VMF in general, and certainly not for EP.

A scalable algorithm to solve DSB for sparse generalized linear models has recently been proposed [4, 14]. It runs orders of magnitude faster than single site updating, and can be used to successfully address MRI design optimization. In order to understand its benefits, we have to view variational inference from a computational standpoint: *which computations are minimally required in order to solve SLM variational inference problems*? In single site updating, each step

[10] Fractional EP is different from the common practice of *damping*, which does not change the fixed points, but does not help either to alleviate problems in this case.

requires marginal mean and variance of $Q(s_j|\boldsymbol{y}) = N(h_j, \sigma^2\rho_j)$, and each potential has to be visited at least once. For large scale SLMs whose model graph is not a tree, each update requires the solution of a $n \times n$ linear system from scratch (see Section 3.1): these algorithms do not provide scalable sparse inference solutions in general. Can we do better with global steps, say by gradient descent? The computation of $\nabla_{\boldsymbol{\gamma}}\phi_{\mathrm{DSB}}$ is dominated by bulk mean and variances computations ("bulk" means that *all* means and variances are required). This is not a problem for the means: all of them are obtained by solving a single $n \times n$ linear system, say by linear conjugate gradients [34]. However, no general method is presently known for approximating all variances in an equally scalable manner, not even if $Q(\boldsymbol{u}|\boldsymbol{y})$ has a sparsely connected graphical model structure such as a nearest-neighbour grid. This fundamental computational difference between bulk mean and variances computation is underlined by theoretical analyses into the convergence of belief propagation in Gaussian Markov random fields [35]. *Any* variational inference algorithm has to compute bulk means and variances eventually, but among different algorithms, those will fare best that require *as few bulk variances computations as possible* until convergence.

Inspired by [25], we can write $\phi_{\mathrm{DSB}}(\boldsymbol{\gamma}) = \log|\boldsymbol{A}| + \phi_{\cup}(\boldsymbol{\gamma})$, where $\phi_{\cup}(\boldsymbol{\gamma})$ is a decoupled function. It is the highly coupled term $\log|\boldsymbol{A}|$ that calls for computing variances. Crucially, $\boldsymbol{\gamma}^{-1} \mapsto \log|\boldsymbol{A}|$ is a concave function, so can be upper bounded by affine functions, making use of Legendre/Fenchel duality once more: $\log|\boldsymbol{A}| \le \boldsymbol{z}^T(\boldsymbol{\gamma}^{-1}) - g^*(\boldsymbol{z})$, $\boldsymbol{z} \succ \boldsymbol{0}$ the normal vector, $g^*(\boldsymbol{z})$ the offset. Naturally, this observation leads to a provably convergent *double loop algorithm*, alternating between outer loop updates (computing $\boldsymbol{z}$, $g^*(\boldsymbol{z})$ so to tangentially fit the affine bound at $\boldsymbol{\gamma}$) and inner loop minimizations of the upper bound $\phi_{\cup}(\boldsymbol{\gamma}) + \boldsymbol{z}^T(\boldsymbol{\gamma}^{-1}) - g^*(\boldsymbol{z})$. In the absence of $\log|\boldsymbol{A}|$, these inner minimizations are solved much more efficiently than $\min_{\boldsymbol{\gamma}\succ\boldsymbol{0}} \phi_{\mathrm{DSB}}$ itself: they are penalized least squares problems of standard form, which can be solved by IRLS, a variant of Newton-Raphson [4]. Variances computations are not required during inner loops at all. For outer loop updates,

$$\boldsymbol{z} \leftarrow \nabla_{\boldsymbol{\gamma}^{-1}} \log|\boldsymbol{A}| = \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T) = \sigma^{-2}(\mathrm{Var}_Q[s_j|\boldsymbol{y}]), \quad g^*(\boldsymbol{z}) \leftarrow \boldsymbol{z}^T(\boldsymbol{\gamma}^{-1}) - \log|\boldsymbol{A}|,$$

requiring the bulk marginal variances computation avoided during the inner loop. Since convergence is typically attained after few such outer loop steps, the double loop algorithm is a scalable DSB solver. It can be applied even if $h(\boldsymbol{\gamma})$ is not convex, see [14] for details.

To sum up, the DSB variational relaxation can be solved for very large models by decoupling the critical $\log|\boldsymbol{A}|$ term by way of a double loop algorithm: most of the work is done in standard form decoupled inner loop optimizations. Can the same idea be applied to VMF or EP? A similar direct approach does not work, because their $h_j^{(\cdot)}$ functions depend on $Q(s_j|\boldsymbol{y})$ and cannot simply be upper bounded. An important point for future research is to find scalable algorithms for these variational approximations.

Finally, is it fair to call the double loop algorithm "scalable"? After all, bulk variances still have to be computed a number of times. Moreover, computing many design score values (3) constitutes a closely related problem. While bulk variances cannot at present be approximated nearly as tractably as solving a linear system, this problem is addressed in numerical mathematics. A general idea is to employ a low rank PCA approximation to $\boldsymbol{A}^{-1}$, featuring only the $k \ll n$ smallest eigenvalues and eigenvectors of $\boldsymbol{A}$. Plugging this in, variances or design scores can be approximated efficiently. The optimality of PCA among all rank-k choices in terms of (co)variance explained is well known. Importantly, typical precision matrices $\boldsymbol{A}$ exhibit a roughly linear[11] spectral decay, so that the k smallest eigenvalues of $\boldsymbol{A}$ can be

[11] The linear spectral decay is *good* news, since Lanczos reveals smallest eigenvalues. It is *bad* news for the overall relative accuracy of variance approximations: their dependence on the interior of the spectrum, never penetrated by Lanczos, is significant.

found efficiently by the Lanczos algorithm [36, 4]. The latter scales superlinearly in k and runs up $O(n\,k)$ memory, so only a limited number of iterations can be done. Proposals exist in numerical mathematics for probing deeper into the spectrum of $\boldsymbol{A}$ [37]. For image models with Markov random field structure, alternative low rank approximations have been proposed that circumvent Lanczos computations [38]. Lastly, variances computation can easily be parallelized. Ultimately, if large scale problems are to be addressed with variational relaxations like DSB *without* accurate bulk variances computation in place, successful solvers will have to be robust to errors committed by estimators such as Lanczos. Empirically, at least for SLMs in the context of MR image data, the double loop algorithm behaves well in that respect [14]. Developing a better understanding of the impact of Gaussian variances errors on variational inference results is an important goal for future research.

3.5. Sparse estimation and sparse inference

We have seen that sparse estimation and sparse inference are different problems. They address different goals (for example, image reconstruction from fixed data, versus design optimization) and come with different algorithmic challenges (for example, bulk Gaussian variances computation is required for variational sparse inference, but not for sparse estimation). Why are they so frequently confused? First, they are based on the same underlying models and prior distributions. Second, in the context of variational approximations, their optimization problems may look deceivingly similar. Recall that SBL is a sparse estimation method, yet is motivated as inference approximation in [11]. By tempering with the $h_j(\gamma_j)$ functions in (5), DSB can be turned into SBL, or into any number of other variants. The highly sparse *estimation* algorithm of [25], solving an instance of SBL, is formally similar to the Laplacian SLM *inference* method of [4]: $h_j(\gamma_j) = \tau_j^2\gamma_j$ in the latter becomes $h_j(\gamma_j) = \log\gamma_j$ in the former: outcomes are drastically different, as are computational demands.

For sparse estimation, the goal is to *eliminate* irrelevant variables. In the context of "SBL-like" methods, this means driving many γ_j to exactly zero. Since $\mathrm{Var}_Q[s_j|\boldsymbol{y}] \le \sigma^2\gamma_j$ (Section 3), $\gamma_j = 0$ leads to $s_j = 0$ almost surely under $Q(\boldsymbol{u}|\boldsymbol{y})$. The latter is a *highly degenerate* distribution, representing correlations and mass only for the small number of non-eliminated variables: certainly not a sensible approximation to the true posterior. Bayesian sequential experimental design based on sparse estimation rather than inference [39] fails for real-world signals [9]: the "posterior" precludes any exploration beyond what has already been extracted from the data, and the procedure gets stuck. Sparse estimation is typically computationally simpler than sparse inference. MAP estimation is a convex quadratic problem for the SLM discussed here, without any approximations. In the double loop algorithm of [25], outer loop updates are simple to do using low rank formulae, since they inherit the exact sparsity of $\boldsymbol{\gamma}$: bulk variances computation is not a hard problem for highly degenerate $Q(\boldsymbol{u}|\boldsymbol{y})$.

Sparse estimation and sparse inference are different problems, but what is their theoretical and algorithmic overlap? Given that sparse estimation receives much more attention, advancing this understanding may lead to benefits for sparse inference. For example, the inner loop problem of the double loop DSB algorithm [4] has the form of a commonly used smooth approximation to MAP estimation for the same model. Variational sparse inference methods have to operate without the computational benefit of exact sparsity in $\boldsymbol{\gamma}$, yet they can still benefit from smoothed estimation technology.

4. Discussion

We have contrasted sparse estimation problems (undersampled reconstruction, denoising, decoding of compressed images) with problems of sparse Bayesian inference (measurement design optimization), where the posterior distribution is approximated beyond locating its mode. We have reviewed common variational relaxations of sparse Bayesian inference, related their criteria and discussed algorithmic differences. Adopting a computation-centered view of these techniques, relating them on the same sparse linear model, we aim to clear up currently prevailing confusions (relying on work by David Wipf and colleagues [40, 22, 25]), to point out major avenues for future research, and to show how sparse estimation and inference can benefit from each other.

Acknowledgments
Some of this work has been done jointly with Hannes Nickisch, Max Planck Institute for Biological Cybernetics, Tübingen. We would like to thank David Wipf and Florian Steinke for helpful discussions. This research is partly funded by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University.

Appendix

In Section 3.2, a variant of variational mean field restricts the maximum in (6) to Gaussian distributions $Q(\boldsymbol{u}|\boldsymbol{y}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $Q(\boldsymbol{u}|\boldsymbol{y})$ is a stationary point, we show that $\boldsymbol{\Sigma}^{-1} = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T\boldsymbol{D}\boldsymbol{B}$ for a diagonal matrix $\boldsymbol{D}$, to that $Q(\boldsymbol{u}|\boldsymbol{y})$ lies in (4), as long as $\boldsymbol{\gamma}$ may have negative entries (if $\boldsymbol{B}$ has full rank n, any mean $\boldsymbol{\mu}$ can be achieved within (4) by varying $\boldsymbol{b}$ alone). Writing $\log P(\boldsymbol{y}|\boldsymbol{u}) = \frac{1}{2}\sigma^{-2}(\boldsymbol{c}^T\boldsymbol{u} - \boldsymbol{u}^T\boldsymbol{P}\boldsymbol{u}) + C$ with $\boldsymbol{P} = \boldsymbol{X}^T\boldsymbol{X}$, and $\nu_j = 2\mathrm{E}_Q[-\log t_j(s_j)]$, then twice the criterion to minimize is $\log|\boldsymbol{\Sigma}| - \sigma^{-2}(\operatorname{tr}\boldsymbol{P}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T) - \boldsymbol{c}^T\boldsymbol{\mu}) - \boldsymbol{1}^T\boldsymbol{\nu}$. Considering the mean $\boldsymbol{\mu}$ to be fixed, the relevant part is $\mathcal{F} = \log|\boldsymbol{\Sigma}| - \sigma^{-2}\operatorname{tr}\boldsymbol{P}\boldsymbol{\Sigma} - \boldsymbol{1}^T\boldsymbol{\nu}$. Here, ν_j depends on $Q(s_j|\boldsymbol{y})$ only, whose means are fixed, and whose variances are $\boldsymbol{\rho} = \operatorname{diag}^{-1}(\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T)$. If $d_j = \partial\nu_j/(\partial\rho_j)$, $\boldsymbol{D} = \operatorname{diag}\boldsymbol{d}$, then $d\mathcal{F} = \operatorname{tr}(\nabla_{\boldsymbol{\Sigma}}\mathcal{F})(d\boldsymbol{\Sigma}) = \operatorname{tr}\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma}) - \sigma^{-2}\operatorname{tr}\boldsymbol{P}(d\boldsymbol{\Sigma}) - \operatorname{tr}\boldsymbol{D}\boldsymbol{B}(d\boldsymbol{\Sigma})\boldsymbol{B}^T$. Solving $\nabla_{\boldsymbol{\Sigma}}\mathcal{F} = \boldsymbol{\Sigma}^{-1} - \sigma^{-2}\boldsymbol{P} - \boldsymbol{B}^T\boldsymbol{D}\boldsymbol{B} = \boldsymbol{0}$ establishes the claim.

In Section 3.3, we claim that moment matching of $\hat{P}_1(\boldsymbol{u})$ is equivalent to minimizing $\mathrm{D}[\hat{P}_1(\boldsymbol{u}) \,\|\, Q(\boldsymbol{u})]$ over all Gaussians $Q(\boldsymbol{u})$. There is a unique Gaussian $Q_1(\boldsymbol{u})$ sharing mean and covariance with $\hat{P}_1(\boldsymbol{u})$, with which $\mathrm{D}[\hat{P}_1(\boldsymbol{u}) \,\|\, Q(\boldsymbol{u})] = \mathrm{D}[\hat{P}_1(\boldsymbol{u}) \,\|\, Q_1(\boldsymbol{u})] + \mathrm{E}_{\hat{P}_1}[\log Q_1(\boldsymbol{u}) - \log Q(\boldsymbol{u})]$. Here, $\log Q_1(\boldsymbol{u}) - \log Q(\boldsymbol{u})$ is a quadratic function in $\boldsymbol{u}$. Since $\hat{P}_1(\boldsymbol{u})$ and $Q_1(\boldsymbol{u})$ have the same moments up to second order, $\mathrm{E}_{\hat{P}_1}[\dots]$ can be replaced by $\mathrm{E}_{Q_1(\boldsymbol{u})}[\dots]$, and the second term becomes $\mathrm{D}[Q_1(\boldsymbol{u}) \,\|\, Q(\boldsymbol{u})] \geq 0$. Therefore, $\mathrm{D}[\hat{P}_1(\boldsymbol{u}) \,\|\, Q(\boldsymbol{u})]$ is minimized uniquely by $Q(\boldsymbol{u}) = Q_1(\boldsymbol{u})$.

Moreover, $\hat{P}_1(\boldsymbol{u}) = \hat{P}_1(s_1)Q_0(\boldsymbol{u}|s_1)$ with $\hat{P}_1(s_1) \propto t_1(s_1)Q_0(s_1)$. By the chain rule of relative entropy, $\mathrm{D}[\hat{P}_1(\boldsymbol{u}) \,\|\, Q(\boldsymbol{u})] = \mathrm{D}[\hat{P}_1(s_1) \,\|\, Q(s_1)] + \mathrm{E}_{\hat{P}_1}[\mathrm{D}[Q_0(\boldsymbol{u}|s_1) \,\|\, Q(\boldsymbol{u}|s_1)]]$, to that $Q_1(\boldsymbol{u}|s_1) = Q_0(\boldsymbol{u}|s_1)$ for the minimizer (zeroing the second term), and the marginal $Q_1(s_1)$ matches moments of $\hat{P}_1(s_1)$, which can be obtained as $Q_1(s_1) \propto e^{\sigma^{-2}(b_j s_j - s_j^2/(2\gamma_j))}Q_0(s_1)$, where b_j, γ_j depend on $Q_0(s_1)$ and $t_1(s_1)$ only.

References

[1] Simoncelli E 1999 *Wavelet Applications in Signal and Image Processing VII: Proc. of SPIE* vol 3813 ed Unser M A, Aldroubi A and Laine A F pp 188–95
[2] Lauterbur P 1973 *Nature* **242** 190–1
[3] Garroway A, Grannell P and Mansfield P 1974 *J. Phys. C: Solid State Phys.* **7** L457–62
[4] Seeger M, Nickisch H, Pohmann R and Schölkopf B 2009 *Advances in Neural Information Processing Systems* vol 21 ed Koller D, Schuurmans D, Bengio Y and Bottou L (MIT Press) pp 1441–8
[5] Donoho D 2006 *IEEE Trans. Inform. Theo.* **52** 1289–306
[6] Candès E, Romberg J and Tao T 2006 *IEEE Trans. Inform. Theo.* **52** 489–509
[7] Weaver J, Xu Y, Healy D and Cromwell L 1991 *Magn. Reson. Med.* **21** 288–95
[8] Lustig M, Donoho D and Pauly J 2007 *Magn. Reson. Med.* **85** 1182–95
[9] Seeger M and Nickisch H 2008 *Int. Conf. on Machine Learning 25* ed McCallum A, Roweis S and Silva R (Omni Press)
[10] Chang H, Weiss Y and Freeman W 2009 Informative sensing Tech. Rep. 0901.4275v1 [cs.IT] ArXiv

[11] Tipping M 2001 *J. Mach. Learn. Res.* **1** 211–44

[12] Seeger M 2008 *J. Mach. Learn. Res.* **9** 759–813

[13] Nickisch H and Seeger M 2009 *Int. Conf. on Machine Learning 26* ed Bottou L and Littman M (Omni Press) pp 761–8

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

[15] Tibshirani R 1996 *J. Roy. Stat. Soc.* B **58** 267–88

[16] Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis* 2nd ed (Springer)

[17] Park T and Casella G 2005 The Bayesian Lasso Tech. rep. University of Florida

[18] Chaloner K and Verdinelli I 1995 *Stat. Sci.* **10** 273–304

[19] Fedorov V 1972 *Theory of Optimal Experiments* (Academic Press)

[20] MacKay D 1991 *Neural Comput.* **4** 589–603

[21] Cover T and Thomas J 1991 *Elements of Information Theory* 1st ed (John Wiley & Sons)

[22] Palmer J, Wipf D, Kreutz-Delgado K and Rao B 2006 *Advances in Neural Information Processing Systems* vol 18 ed Weiss Y, Schölkopf B and Platt J (MIT Press)

[23] Boyd S and Vandenberghe L 2002 *Convex Optimization* (Cambridge University Press)

[24] Jaakkola T and Jordan M 2000 *Stat. Comput.* **10** 25–37

[25] Wipf D and Nagarajan S 2008 *Advances in Neural Information Processing Systems* vol 20 ed Platt J, Koller D, Singer Y and Roweis S (MIT Press)

[26] Girolami M 2001 *Neural Comput.* **13** 2517–32

[27] Wainwright M J and Jordan M I 2008 *Found. Trends Mach. Learn.* **1** 1–305

[28] Minka T 2001 *Uncertainty in Artificial Intelligence 17* ed Breese J and Koller D (Morgan Kaufmann)

[29] Opper M and Winther O 2000 *Neural Comput.* **12** 2655–84

[30] Kushner H and Budhiraja A 2000 *IEEE Trans. Automat. Contr.* **45** 580–5

[31] Opper M and Winther O 2005 *J. Mach. Learn. Res.* **6** 2177–204

[32] Minka T 2004 Power EP Tech. rep. Microsoft Research, Cambridge

[33] Bogachev V 1998 *Gaussian Measures* Mathematical Surveys and Monographs (American Mathematical Society)

[34] Golub G and Van Loan C 1996 *Matrix Computations* 3rd ed (Johns Hopkins University Press)

[35] Malioutov D, Johnson J and Willsky A 2006 *J. Mach. Learn. Res.* **7** 2031–64

[36] Schneider M and Willsky A 2001 *SIAM J. Sci. Comput.* **22** 1840–64

[37] Bekas C, Kokiopoulou E and Saad Y 2008 *SIAM J. Mat. Annal. Appl.* **30** 397–418

[38] Malioutov D, Johnson J, Choi M and Willsky A 2008 *IEEE Trans. Signal Proces.* **56** 4621–34

[39] Ji S and Carin L 2007 *Int. Conf. on Machine Learning 24* ed Ghahramani Z (Omni Press)

[40] Wipf D, Palmer J and Rao B 2004 *Advances in Neural Information Processing Systems* vol 16 ed Thrun S, Saul L and Schölkopf B (MIT Press)