Gaussian Covariance and Scalable Variational Inference

Matthias W. Seeger MSEEGER@MMCI.UNI-SAARLAND.DE
Saarland University and Max Planck Institute for Informatics, Campus E1.7, 66123 Saarbruecken, Germany

Abstract

We analyze computational aspects of variational approximate inference techniques for sparse linear models, which have to be understood to allow for large scale applications. Gaussian *covariances* play a key role, whose approximation is computationally hard. While most previous methods gain scalability by not even representing most posterior dependencies, harmful factorization assumptions can be avoided by employing data-dependent low-rank approximations instead. We provide theoretical and empirical insights into algorithmic and statistical consequences of low-rank covariance approximation errors on decision outcomes in nonlinear sequential Bayesian experimental design.

1. Introduction

Sparse linear models (SLMs) enjoy enormous popularity in high-dimensional statistics, signal and image processing, and machine learning. A large part of this success story is due to regard for computational details: *maximum a posteriori* (MAP) estimation is formulated in terms of convex problems, which are reduced to standard primitives of numerical mathematics and digital signal processing. In such point estimation techniques, the Bayesian posterior is used as a criterion to be maximized rather than a distribution to be approximated and queried. Many applications require posterior information beyond its mode's location. Decision theory and Bayesian experimental design can be used to optimize sampling patterns (Seeger et al., 2009) or data acquisition, and sparse *bilinear* model reconstruction is greatly improved by Bayesian averaging (Levin et al., 2009). However, today's approximate *inference* technology lags far behind MAP estimation in terms of scalability, robustness, and theoretical understanding.

In this paper, we focus on computational aspects of *scalable variational inference* for large SLMs. Bayesian inference is hard and useful for the same underlying reason: the emergence of very many nonlocal *dependencies* in the posterior distribution. In the large scale continuous-variable context, these are approximated by *Gaussian covariances* of restricted structure and dimensionality. The choice of these restrictions not only impacts the final best fit to the posterior, but also the optimization process leading there. By far most methods to date attain scalability through factorization assumptions, whereby all dependencies are forced into a predetermined form, and most of them are ruled out up front. In contrast, Seeger et al. (2009) show how to avoid factorizations entirely, using low-rank covariance approximations such as PCA or the Lanczos algorithm (Schneider & Willsky, 2001) instead. The latter concept of tracking a limited number of principal covariance directions alongside the variational optimization has advantages in practice, since most Bayesian decision making or experimental design applications are driven by dominating modes of posterior dependencies.

We point out the fundamental role of Gaussian (co)variance computations for large scale variational inference and experimental design in Section 2, and review approximation methods in Section 3. Our main contribution is an analysis of how low-rank Gaussian covariance approximations affect inference outcomes in the framework of Seeger et al. (2009). First, we prove that if covariances are approximated by PCA rather than computed exactly, their algorithm remains convergent. Our argument is based on convexity of spectral functions (Davis, 1957). Second, we show that in the context of SLM inference, PCA approximation errors lead to a systematic *strengthening* of the sparsity regularization. A running example in this paper is optimizing real-world image acquisition (adaptive compressive sensing) by Bayesian experimental design (Seeger et al., 2009). We discuss the impact of PCA approximations on design score computations and sequential decisions, and provide experimental results on

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

real-world images in Section 5.

2. Variational Inference for Sparse Linear Models

We are interested in *sparse linear models* (SLMs) applied to image reconstruction (see (Seeger, 2008) for a detailed exposition). A latent image $\boldsymbol{u} \in \mathbb{R}^n$ (n pixels) is sampled by way of a design matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$. Observations $\boldsymbol{y} \in \mathbb{R}^m$ are modelled as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ is noise of variance σ^2. For example, $\boldsymbol{X}$ is a partial discrete Fourier transform in MRI reconstruction applications (Seeger et al., 2009). The sparsity of filter coefficients $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u} \in \mathbb{R}^q$ (such as wavelet coefficients or spatial derivatives) is encouraged by way of a Laplacian prior distribution $P(\boldsymbol{u}) \propto \prod_{j=1}^q e^{-\tau_j |s_j/\sigma|}$. We are interested in *sparse inference*, for example to approximate the posterior covariance in order to optimize the design $\boldsymbol{X}$ (see Section 2.2). Even for modest image resolutions, n (number pixels) is beyond 100000, with m (number samples) and q (number sparsity coefficients) of the same order. Note that in many image reconstruction applications, $\boldsymbol{X}$ is neither sparse, nor has simple graphical model structure.

The posterior distribution has the form

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2 \boldsymbol{I}) \prod\nolimits_{j=1}^q e^{-\tau_j |s_j/\sigma|}, \qquad (1)$$

$\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$, which we would like to integrate against and obtain moments of. This is hard for two reasons coming together. First, $P(\boldsymbol{u}|\boldsymbol{y})$ is highly *coupled*, since $\boldsymbol{X}$ is neither diagonal nor sparse. Second, it is *non-Gaussian* due to the Laplacian prior potentials. In large scale regimes, a third problem is the sheer size even of basic moments such as the covariance matrix. In cases of interest here, inference is practically intractable even for linear models with Gaussian prior.

2.1. Scalable Algorithms

How can $P(\boldsymbol{u}|\boldsymbol{y})$ be approximated at large scales? Most methods make use of a *Gaussian* approximation $Q(\boldsymbol{u}|\boldsymbol{y}; \boldsymbol{\gamma})$, either fitting Q to P globally, or using Q as carrier for self-consistency equations between marginals. The rationale is that global covariances can still be represented this way, while Gaussian integrals are tractable to compute. In this section, we show that large scale variational inference crucially relies on bulk computation of Gaussian variances.

Most algorithms to date can be grouped into two classes: either [1] $Q(\boldsymbol{u}|\boldsymbol{y})$ is restricted to factorize, with factors ranging over small disjoint subsets of $\boldsymbol{u}$, or [2] updates are done based on marginals $Q(s_j|\boldsymbol{y})$, $j = 1, \dots, q$, kept up-to-date by message passing. Both notions lead to easily implementable algorithms, iterating between local factor or node updates and Gaussian message passing. Unfortunately, none of these approaches result in scalable algorithms in general. Both factorization assumptions and single-marginal updating lead to non-convex inference relaxations in all cases we know of. More important, while each single update is easy to do, far too many of them are required until convergence. For [2], we require q updates to even visit each marginal, and the absence of a sparsity or graph structure of $\boldsymbol{X}$ precludes fast message passing in between: an $n \times n$ linear system has to be solved for each update. In order to be applied at very large scales, a variational algorithm has to do few *global* iterations[1] until convergence, which in turn have to be reducible to scalable computational problems.

A scalable variational inference method has been proposed by Seeger et al. (2009), see also (Nickisch & Seeger, 2009). We sketch some details to outline computational demands and prepare the ground for further analysis. Variational methods target the log partition function $\log Z$ of (1), the cumulant-generating function of the posterior. It is lower bounded by plugging in the representation $e^{-\tau_j |s_j/\sigma|} = \max_{\gamma_j > 0} e^{-(s_j/\sigma)^2/(2\gamma_j) - \tau_j^2 \gamma_j/2}$ for Laplacian sites, then interchanging $\max_{\boldsymbol{\gamma} \succ \boldsymbol{0}}$ with the integral over $\boldsymbol{u}$. The variational problem constitutes finding the closest bound over variational parameters $\boldsymbol{\gamma}$: $\min_{\boldsymbol{\gamma} \succ \boldsymbol{0}} \phi(\boldsymbol{\gamma})$ with

$$\begin{aligned} \phi(\boldsymbol{\gamma}) &= \log |\boldsymbol{A}| + (\boldsymbol{\tau}^2)^T \boldsymbol{\gamma} + \min_{\boldsymbol{u}} R(\boldsymbol{u}, \boldsymbol{\gamma}), \\ R &:= \sigma^{-2} \left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s} \right), \end{aligned} \qquad (2)$$

where $\boldsymbol{\Gamma} := \operatorname{diag} \boldsymbol{\gamma}$. The posterior $P(\boldsymbol{u}|\boldsymbol{y})$ is fitted by a *Gaussian* approximation $Q(\boldsymbol{u}|\boldsymbol{y}) = N(\boldsymbol{u}|\boldsymbol{u}_*, \sigma^2 \boldsymbol{A}^{-1})$ parameterized in terms of $\boldsymbol{\gamma}$, in that $\boldsymbol{A} = \boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{B}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{B}$ and $\boldsymbol{u}_* = \boldsymbol{u}_*(\boldsymbol{\gamma})$. Denote posterior marginals by $Q(s_j|\boldsymbol{y}) = N(h_j, \sigma^2 \rho_j)$ in the sequel, $\boldsymbol{\rho} = (\rho_j)$. It is easy to see that $\rho_j = \sigma^{-2} \mathrm{Var}_Q[s_j|\boldsymbol{y}] \le \gamma_j$: the variational parameters directly control the variances. SLMs implement *selective shrinkage*, in that most $|s_j|$ are strongly forced to small values, while some (the "relevant" ones) are shrunken little at all. The key statistic for sorting coefficients in this way is *posterior variance*, and the role of the γ_j is to implement selective shrinkage within the Gaussian approximation $Q(\boldsymbol{u}|\boldsymbol{y})$.

[1] In principle, methods from [1] and [2] can be run doing parallel updates, which would look like "global" steps. However, single components of such parallel updates are derived by assuming the rest of $Q(\boldsymbol{u}|\boldsymbol{y})$ remains static: they are not coupled themselves. To our knowledge, parallel updating for [1] or [2] has not been reported to run faster for SLMs than simpler sequential variants.

The variances are not only statistically decisive, but also computationally. For gradient-based minimization of ϕ, we certainly require $\nabla_{\boldsymbol{\gamma}^{-1}} \log|\boldsymbol{A}| = \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T) = \boldsymbol{\rho}$. While all marginal *means* (h_j) can be computed solving a single linear system by conjugate gradients (Golub & Van Loan, 1996), bulk *variances* computation is much more difficult. A key property of the scalable algorithm is that Gaussian variances have to be computed few times only until convergence. By affinely upper-bounding the coupling term $\log|\boldsymbol{A}|$ in (2), we iterate between the following two steps (called *outer loop update* and *inner loop minimization*):

$$\begin{aligned}
&\boldsymbol{z} \leftarrow \boldsymbol{\rho} = \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T),\\
&\boldsymbol{u}_* \leftarrow \operatorname*{argmin}_{\boldsymbol{u}} \phi_{\boldsymbol{z}},\ \gamma_j \leftarrow \frac{1}{\tau_j}\sqrt{z_j + \left|\frac{s_{*,j}}{\sigma}\right|^2},\ \boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*,\\
&\phi_{\boldsymbol{z}} = \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + 2\sum\nolimits_j \tau_j\sqrt{z_j + (|s_j|/\sigma)^2}.
\end{aligned} \tag{3}$$

Updating $(\boldsymbol{u}_*, \boldsymbol{\gamma})$ is a standard penalized least squares problem, which can be solved at large scales. Variance computations are required only for computing $\boldsymbol{z}$, which happens few times until convergence.

To sum up, the main computational primitives of large scale SLM variational inference are *Gaussian means and variances*, the latter are much more difficult to approximate. Gaussian mean computations (least squares) are bread-and-butter in any computational discipline, while variance computations, not required for point estimation, are less frequently addressed. Among inference algorithms aiming at large scales, those stand out which require Gaussian variances as seldomly as possible. Note that variances are not required in MAP and most other sparse estimation methods. In the context of large scale sparse linear models, the requirement of Gaussian variances is the most important computational difference between variational *sparse inference* and *sparse estimation.* Since they can in general not be approximated to close relative accuracy, it is important to understand effects of Gaussian variance errors on variational inference outcomes.

2.2. Sequential Bayesian Experimental Design

Approximating a SLM posterior $P(\boldsymbol{u}|\boldsymbol{y})$ by a Gaussian $Q(\boldsymbol{u}|\boldsymbol{y})$ is hard at large scales. But how accurate do we have to be? Which additional structural restrictions on Q can be tolerated? Fortunately, while a uniformly close approximation of $P(\boldsymbol{u}|\boldsymbol{y})$ or its marginals $P(s_j|\boldsymbol{y})$ is presently unattainable at large scales, this is too much to ask for in typical applications, such as the image acquisition optimization problem. In this section, we argue that in many such decision making scenarios, the critical information are the *maximum* covariance directions of the posterior, a fact which directly motivates PCA and Lanczos covariance approximations discussed in Section 3.

In Bayesian experimental design (ED), the design matrix $\boldsymbol{X}$ is optimized sequentially, appending parts $\boldsymbol{X}_*$ which maximize an information gain score, in this case $\Delta(\boldsymbol{X}_*) := \log|\boldsymbol{I} + \boldsymbol{X}_*\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}]\boldsymbol{X}_*^T|$ (Seeger et al., 2009). Considering $\{\Delta(\boldsymbol{X}_*)\}$ over many candidates $\boldsymbol{X}_*$, it is the posterior covariance $\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}]$ these score values depend upon: the mode or mean of $Q(\boldsymbol{u}|\boldsymbol{y})$ are irrelevant. $\Delta(\boldsymbol{X}_*)$ measures the overlap of $\boldsymbol{X}_*$ with leading eigendirections of $\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}]$, the directions of *maximum posterior covariance.* In order to drive Bayesian ED successfully, it is not necessary to closely approximate all of $P(\boldsymbol{u}|\boldsymbol{y})$: decision making depends mainly on its leading covariance eigendirections. Consider the extreme case, where $\boldsymbol{X}_*$ is a single row and all unit norm vectors are candidates: the score maximizer is the single maximal eigenvector of $\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}] = \sigma^2\boldsymbol{A}^{-1}$. While all of $\boldsymbol{A}$ cannot even be stored, let alone inverted, its leading eigenvector can be obtained tractably (see Section 3). On the other hand, it is not enough to approximate posterior marginals only, or to fit any factorized distribution to $P(\boldsymbol{u}|\boldsymbol{y})$: neither give sufficient information about the leading covariance eigenmodes in general.

3. Approximating Gaussian Variances

We exposed Gaussian variances computation as general bottleneck for scalable variational inference, no matter which specific method is used. As discussed in Section 2.1, most previous algorithms are not scalable up front, since variances are computed one by one, and consequently there has been little machine learning interest in Gaussian variances so far. Still, their dominant role in (Seeger et al., 2009) and scalable inference in general motivates closer attention.

Methods have been proposed in spatial statistics (Willsky, 2002), often exploiting graph structure or sparsity of $\boldsymbol{X}$, neither of which is present in our case. A general idea for approximating variances is to estimate them from a number $L \ll n$ of linear projections, thus to approximate $\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}]$ by a low-rank matrix. In special cases, good projections can be constructed based on prior knowledge about signal structure (Malioutov et al., 2008). Perhaps the most promising general approach is to choose projections that capture as much covariance as possible: the L *principal components.* Namely, if $\boldsymbol{A} \approx \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ (L smallest eigenvalues/-vectors), then $\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}] \approx \sigma^2 \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T$ and $\boldsymbol{\rho} \approx \boldsymbol{\rho}^{(L)} := \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T\boldsymbol{B}^T)$. Of course, L-PCA remains an academic exercise without a scalable way of computing $\boldsymbol{U}$, $\boldsymbol{\Lambda}$ for L sufficiently large. Fortunately, this can often be done by way of the Lanczos algorithm (Schneider & Willsky, 2001).

Details about the Lanczos algorithm are found in (Golub & Van Loan, 1996). At one matrix-vector multiplication (MVM) with $\boldsymbol{A}$ per iteration, we obtain $\boldsymbol{Q}_k^T\boldsymbol{A}\boldsymbol{Q}_k = \boldsymbol{T}_k$, $\boldsymbol{Q}_k \in \mathbb{R}^{n\times k}$ orthonormal, $\boldsymbol{T}_k$ tridiagonal. $\boldsymbol{A}$'s extremal eigenvalues/-vectors are closely approximated by those of $\boldsymbol{Q}_k\boldsymbol{T}_k\boldsymbol{Q}_k^T$ (the latter are called Ritz values/vectors). Eigenconvergence can easily be monitored inside the method: the SVD of $\boldsymbol{T}_k$ can be done in $O(k^2)$. For example, the maximum eigenvector of $\boldsymbol{A}$ is typically obtained to high accuracy after few iterations (see Section 2.2). In practice, L-PCA is approximated by running Lanczos for $K > L$ steps, until the L smallest eigenvalues of $\boldsymbol{A}$ have converged. Whether this is feasible or not, depends on the spectral structure of $\boldsymbol{A}$. Lanczos convergence theory (Golub & Van Loan, 1996) states that eigenvalues converge from the fringe of the spectrum inwards, roughly ordered by the gap size between neighbouring entries. For example, if $\mathrm{spec}(\boldsymbol{A})$ decreases geometrically, eigenvalues converge from largest to smallest. However, precision matrices in typical SLM scenarios show roughly linear spectral decay (Seeger, 2009), so that largest *and* smallest eigenvalues converge even for $K \ll n$. At present, it may well be the best general method for variance approximations in the context of SLM variational inference. Of course, rather than settling for *L-PCA* after K Lanczos steps, we may as well make use of the complete Lanczos representation, approximating $\boldsymbol{A}^{-1}$ by $\boldsymbol{Q}_k\boldsymbol{T}_k^{-1}\boldsymbol{Q}_k^T$ instead of $\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T$, referred to as *K-Lanczos* approximation in the sequel. The corresponding variance estimator is $\boldsymbol{\rho}_k = \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{Q}_k\boldsymbol{T}_k^{-1}\boldsymbol{Q}_k^T\boldsymbol{B}^T)$.

The Lanczos algorithm is not easy to implement or analyze, and comes with a higher cost than conjugate gradients. Ironically, the convergence of Ritz values causes the difficulties: they continue to contaminate subsequent steps by way of numerical round-off, avoided only by orthogonalizing each new vector against all previously converged Ritz vectors (or all columns of $\boldsymbol{Q}_k$). Therefore, $\boldsymbol{Q}_k$ has to be stored at $O(n\,K)$, and orthogonalization costs up to $O(n\,K^2)$. Can accurate variances be obtained with few iterations? In general, this is possible only if $\boldsymbol{A}$'s spectrum decays geometrically, which does not happen for typical system matrices in our case. The part of the spectrum we miss out on with moderate K carries substantial mass. $\boldsymbol{A}^{-1}$ is not closely approximated by *any* matrix of low rank in terms of (co)variance explained, and relative errors of $\boldsymbol{\rho}_k$ tend to be substantial. This uniformly bleak picture will be constrasted with a fine-grained analysis in Section 4.1, exposing structure in Lanczos variance errors which can be beneficial for SLM inference applications.

Can the Lanczos algorithm be improved in the context of variance approximation? After all, we do not require eigenvectors/-values as such, but a specific estimate based on them only. First, reorthogonalization cannot be skipped. Doing so renders the Lanczos algorithm practically useless if more than the leading eigendirection is required. Deflation time can be saved by selective orthogonalization (Golub & Van Loan, 1996), whereby Ritz vector convergence is monitored during the course of the algorithm. Unfortunately, this requires even more memory, and in some comparisons of ours did not lead to substantial speed-ups.

4. Consequences of Lanczos/PCA Approximations

In this section, we analyze effects of Lanczos variance approximation errors on SLM variational inference, within the double loop algorithm of Seeger et al. (2009) sketched in (3). First, we highlight the overall statistical role played by structures in these errors. Second, by using a result on convexity of spectral functions, we show that the convergence proof of the algorithm is retained with L-PCA variance approximation.

4.1. Lanczos Approximations and Sparsity

Gaussian variances are fundamental for sparse Bayesian inference (Section 2.1) and experimental design (Section 2.2), yet cannot be obtained to high relative accuracy for large scale models of interest (Section 3). Why can we still obtain sensible results? In this section, we aim to understand the *statistical* role of Lanczos variance approximation errors. Recalling from Section 2.1 that sparsity priors enforce selective shrinkage, we show that this effect is strengthened by variance errors.

First, both $\boldsymbol{\rho}^{(L)}$ (L-PCA) and $\boldsymbol{\rho}_K$ (K-Lanczos) are monotonically nondecreasing (w.r.t. L, K) and lower-bound $\boldsymbol{\rho}$ in each component. This is immediate for $\boldsymbol{\rho}^{(L)}$ (since eigenvalues are positive) and easy to show for $\boldsymbol{\rho}_K$. Interestingly, the ratio of underestimation $\rho_{K,j}/\rho_j$ has a clear structure. In (Seeger, 2009), $\rho_j \mapsto \rho_{K,j}/\rho_j$ are plotted for different values of K. The error is smallest for those coefficients whose true variance ρ_j is large, while coefficients with moderate true ρ_j are most strongly damped. As K grows, these worst ratios are lifted towards 1, while the errors for the largest ρ_j, smallish to begin with, are affected least. In summary, there is a stable *structure* in the Lanczos variance errors, coming from the algorithm's working. The largest eigenvalues of $\boldsymbol{A}^{-1}$ converge rapidly, even for small K. In general, the largest ρ_j depend most strongly on these largest eigenvalue contributions.

Still, we run inner loop optimizations, plugging in values for $\boldsymbol{z}$ which are overall substantially too small (since $\boldsymbol{z} \leftarrow \boldsymbol{\rho}$ is replaced by $\boldsymbol{z} \leftarrow \boldsymbol{\rho}_K$). In (3), the z_j feature in the penalty term $(z_j + |s_j/\sigma|^2)^{1/2}$, whose strength (in terms of enforcing $|s_j| \to 0$) grows with shrinking[2] z_j. *The effect of Lanczos variance errors within this framework is to strengthen soft sparsity penalization.* This does not happen uniformly across s_j: if the true $z_j \leftarrow \rho_j$ is among the largest coefficients, shrinkage by Lanczos is least pronounced, while moderately small true z_j are strongly diminished. The selective shrinkage effect discussed in Section 2.1 is strengthened by Lanczos variance errors. Those coefficients most relevant under exact computation are least affected by Lanczos variance errors, but the damping of less relevant ones is amplified. While there is no proof that Lanczos errors do not hurt SLM variational inference in general, they do not work against this decisive effect.

4.2. Convergence of Double-Loop Algorithm

Recall from Section 2.1 that the inference relaxation of interest here is a convex optimization problem for Laplacian potentials, and its double loop algorithm is provably convergent in general (Seeger et al., 2009). These statements hold if Gaussian variances are computed exactly, which is impossible at large scales. It is natural to ask which of these beneficial properties are provably *retained* if variances are approximated by L-PCA. Similar questions should be asked more frequently in machine learning. It is often the case that desirable properties like convexity or guaranteed convergence are proven assuming exact computations, yet real-world experiments are run based on low-rank or subsampling approximations. In all cases we are aware of, the effects of such approximations on former properties remain unanalyzed.

In order to lift "exact" properties for methods with embedded approximations, the challenge is to characterize the latter in a strong way, so that even with the approximation in place, crucial steps in the "exact computation" proof remain valid. This is possible for L-PCA, using a result on convexity of spectral functions (Davis, 1957), as summarized in the following theorem.

Theorem 1 *Consider the tractable variant of the method of (Seeger et al., 2009), replacing outer loop updates in (3) by $\boldsymbol{z} \leftarrow \boldsymbol{\rho}^{(L)}$ (L-PCA variance approximation). This modified algorithm is provably convergent, in the general setting given by Nickisch & Seeger (2009). However, the convexity of the modified relaxation may be compromised.*

Proof. Inspecting the convergence proof in (Seeger et al., 2009), crucial points are the concavity of $\boldsymbol{\gamma}^{-1} \mapsto \log|\boldsymbol{A}|$, and that $\boldsymbol{z}$ is updated as $\nabla_{\boldsymbol{\gamma}^{-1}} \log|\boldsymbol{A}|$. This ensures that ϕ and the inner loop criterion $\phi_{\boldsymbol{z}}$ meet tangentially at current points $\boldsymbol{\gamma}$, see also (Wipf & Nagarajan, 2008). For L-PCA approximations $\boldsymbol{A} \approx \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ ($\boldsymbol{U} \in \mathbb{R}^{n\times L}$ orthonormal, $\boldsymbol{\Lambda}$ the L smallest eigenvalues), $\log|\boldsymbol{A}|$ is replaced by $\log|\boldsymbol{\Lambda}|$, and $\boldsymbol{\rho}$ by $\boldsymbol{\rho}^{(L)} = \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T\boldsymbol{B}^T)$, where $\boldsymbol{A}$, $\boldsymbol{U}$, $\boldsymbol{\Lambda}$ are mappings of $\boldsymbol{\gamma} \succ \boldsymbol{0}$. We have to show that $\boldsymbol{\gamma}^{-1} \mapsto \log|\boldsymbol{\Lambda}|$ is concave, and that $\nabla_{\boldsymbol{\gamma}^{-1}} \log|\boldsymbol{\Lambda}| = \boldsymbol{\rho}^{(L)}$. We assume that at each $\boldsymbol{\gamma} \succ \boldsymbol{0}$ of interest, the smallest L eigenvalues of $\boldsymbol{A}$ are separated (by continuity, this holds in a small environment as well), so that eigenvalue derivatives are well-defined. If λ is an eigenvalue at $\boldsymbol{\gamma}$ with unit eigenvector $\boldsymbol{u}$, then $(d\boldsymbol{u})^T\boldsymbol{u} = 0$ (since $\boldsymbol{u}^T\boldsymbol{u} = 1$ around $\boldsymbol{\gamma}$), thus $d\lambda = \boldsymbol{u}^T(d\boldsymbol{A})\boldsymbol{u} + 2\boldsymbol{u}^T\boldsymbol{A}(d\boldsymbol{u}) = \boldsymbol{u}^T(d\boldsymbol{A})\boldsymbol{u} + 2\lambda\boldsymbol{u}^T(d\boldsymbol{u}) = \boldsymbol{u}^T(d\boldsymbol{A})\boldsymbol{u}$. Therefore, $\nabla_{\boldsymbol{\gamma}^{-1}} \log|\boldsymbol{\Lambda}| = \sum_{i>n-L} \lambda_i^{-1}(\boldsymbol{B}\boldsymbol{u}_i)^2 = \boldsymbol{\rho}^{(L)}$.

Proving the concavity of $\boldsymbol{\gamma}^{-1} \mapsto \log|\boldsymbol{\Lambda}|$ is harder, due to the implicit definition of $\boldsymbol{\Lambda}$. We draw on results for *spectral functions*. Such $f(\boldsymbol{A})$ are induced from symmetric functions $f : \mathbb{R}^n \to \mathbb{R}$ ($f(\boldsymbol{P}\boldsymbol{x}) = f(\boldsymbol{x})$ for any coefficient permutation $\boldsymbol{P} \in \mathcal{P}_n$) by way of $f(\boldsymbol{A}) := f(\mathrm{spec}(\boldsymbol{A}))$, where $\mathrm{spec}(\boldsymbol{A})$ are the eigenvalues of $\boldsymbol{A}$. Clearly, $\boldsymbol{A} \mapsto \log|\boldsymbol{\Lambda}| = \sum_{i>n-L} \lambda_i(\boldsymbol{A})$ is a spectral function. It is shown in (Davis, 1957) that if f is symmetric, convex, and lower semicontinuous, then its induced spectral function is convex and lower semicontinuous over Hermitian matrices. Let $h(\boldsymbol{x}) := -\sum_{i=1}^{L} \log(x_i)$ for $\boldsymbol{x} \succ \boldsymbol{0}$, $h(\boldsymbol{x}) := \infty$ elsewhere, and $f(\boldsymbol{x}) := \max_{\boldsymbol{P}\in\mathcal{P}_n} h(\boldsymbol{P}\boldsymbol{x})$. Since $-\log$ is convex and decreasing, so is h. f is convex as maximum over convex functions, and since h is decreasing, it is the L smallest entries of $\boldsymbol{x}$ that $f(\boldsymbol{x})$ depends on: $f(\boldsymbol{A}) = -\log|\boldsymbol{\Lambda}|$. Therefore, the modified algorithm remains provably convergent.

The convexity of $\phi(\boldsymbol{\gamma})$ hinges on the convexity of $\boldsymbol{\gamma} \mapsto \log|\boldsymbol{A}|$ (Nickisch & Seeger, 2009). However, $\boldsymbol{\gamma} \mapsto \log|\boldsymbol{\Lambda}|$ is not convex in general, as the following counter-example shows: $n = q = 2$, $L = 1$,

[2] If $z_j = 0$ throughout, this becomes $|s_j|/\sigma$, and the inner loop minimization reduces to MAP estimation.

$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_2$, $\boldsymbol{B} = \boldsymbol{I}_2$, so that $\log|\boldsymbol{A}| = \log|\boldsymbol{I}_2 + \boldsymbol{\Gamma}^{-1}|$. Then, $\log\lambda_2 = \min_i \log(1 + 1/\gamma_i)$, convex at each $\boldsymbol{\gamma}$ with $\gamma_1 \neq \gamma_2$. But for $\boldsymbol{\gamma} = (1\,1)^T + t\boldsymbol{p}$, $f(t) = \min_i \log(1 + 1/(1 + p_i t))$ is not convex at $t = 0$ if $p_1 \neq p_2$. If $0 < p_1 < p_2$, the argmin is 2 for $t > 0$, 1 for $t < 0$. The derivative of a component at $t = 0$ is $-p_i/2$, so that $f'(t) \to -p_1/2$ for $t \to 0$, $t < 0$, and $f'(t) \to -p_2/2$ for $t \to 0$, $t > 0$. Since $-p_1/2 > -p_2/2$, $f(t)$ is not convex at $t = 0$. Note that $\boldsymbol{\gamma} \mapsto \log|\boldsymbol{\Lambda}|$ *is* locally convex in regions where eigenvalues do not cross over. However, it fails to be globally convex in general. This concludes the proof.

Even if variances are approximated by L-PCA with any $L \geq 1$, the double loop algorithm is guaranteed to converge. Of course, we optimize the wrong criterion in this case, but the method is self-consistent and enjoys the same convergence properties. Our proof does not extend to K-Lanczos variance approximations. If $\log|\boldsymbol{A}|$ is replaced by $\log|\boldsymbol{T}_K|$, then $\nabla_{\boldsymbol{\gamma}^{-1}} \log|\boldsymbol{T}_K| = \boldsymbol{\rho}_K$. However, we are lacking a description of $\boldsymbol{\gamma} \mapsto \boldsymbol{T}_K$ strong enough in order to prove (or disprove) global concavity of $\boldsymbol{\gamma}^{-1} \mapsto \log|\boldsymbol{T}_K|$. Finally, the concavity of $\boldsymbol{A} \mapsto \log|\boldsymbol{\Lambda}|$ may be more generally useful for analyzing PCA approximations embedded in machine learning methods. For example, there is recent interest in estimating Gaussian model structure by way of penalized maximum likelihood with l_1 potentials on the entries of the precision matrix $\boldsymbol{P}$ (Banerjee et al., 2008). The likelihood part of the criterion has the form $\operatorname{tr}\boldsymbol{E}^T\boldsymbol{P} - \log|\boldsymbol{P}|$: a convex function in $\boldsymbol{P}$ or some linear parameterization. If $\boldsymbol{P}$ is large, a natural approximation would be L-PCA. In this case, our result above implies that the problem remains convex if $\log|\boldsymbol{P}|$ is replaced by $\log|\boldsymbol{\Lambda}|$, $\boldsymbol{\Lambda}$ the L smallest eigenvalues of $\boldsymbol{P}$.

4.3. Evaluations of Approximate Inference

Arguments in this section are developed for SLM variational inference with the double loop algorithm discussed above, but may have wider significance. Besides computational tractability, it is important to understand differences in *robustness* to variance errors across algorithms: does the double loop algorithm stand out, or are other methods (see Section 2.1) equally tolerant (while much slower)? In our view, it does not make much sense to relate approximate Bayesian techniques to the inherently intractable ideas they tend to be motivated with, or to grant too much value to such a motivation in the first place. After all, inference is used to drive real-world problems, and approximation errors have to understand in these contexts (see Section 2.2). Moreover, variational inference methods should be compared against today's feasible alternatives for these decision making problems. For example, switching to a model of simpler graphical structure, we risk to optimize acquisitions for the wrong reconstruction setup, thus to solve the wrong problem more accurately. As noted in Section 2.1, we can opt for variational mean field Bayes (Attias, 2000), tractable at large scales only with many factorization assumptions. However, most covariances are not represented in a factorized $Q(\boldsymbol{u}|\boldsymbol{y})$. The choice of which dependencies to suppress is typically done beforehand, without even looking at data. Factorization assumptions seem disadvantageous in decision scenarios like experimental design, where the dominant posterior covariances are all that matters. While more difficult to implement, Lanczos (PCA) approximations provide a superior alternative in this context, since leading posterior covariances can be tracked in a data-dependent manner.

Our arguments have implications for how (variational) approximate inference methods should be evaluated. At present, the quality of a new technique is evaluated by comparing the relative accuracy of marginals on small regular graphs, where brute force exact computations can still be done. But virtually no applications of approximate Bayesian inference rely crucially on highly accurate marginal numbers. Typically, decision scenarios such as in Section 2.2 are faced: what matters is that the highest scoring candidates come out on top for the approximation as well. When Lanczos approximations are used for moderate K, almost all variances are underestimated substantially (Section 3), but special structure in these errors implies that they do not adversely affect selective shrinkage in SLM inference (Section 4.1). Similar to marginal variances, Lanczos approximations of design score curves $\Delta(\boldsymbol{X}_*)$ are globally much too small, but its maximum points tend to stick out even for small K (see Section 2.2). There are many other practically relevant properties of a method, such as computational efficiency, robustness, uniqueness, guaranteed convergence, ease of use, and some of these may well be more important than accuracy of marginals. A main message of this paper is that variational inference techniques should be analyzed and compared on real-world decision scenarios. Testing them on artificial problems may paint a misleading picture.

5. Experiments

Recall the image acquisition optimization setup from Section 2.2. The underlying assumption is that with good posterior approximations, highly informative designs will be found (taylored to the reconstruction model), and MAP reconstruction results on a test set will be improved. The goal of our study[3] is to test how reconstruction quality varies with different levels of Gaussian variance errors. In line with Section 4.3, we choose a realistic setup: image reconstruction from noisy Fourier measurements (required for magnetic resonance imaging). We use the model setup previously employed in (Seeger et al., 2009; Seeger & Nickisch, 2008): $\boldsymbol{B}$ consists of an orthonormal wavelet transform $\boldsymbol{B}_a$ and horizontal/vertical differences $\boldsymbol{B}_r$, corresponding prior parameters are τ_a, τ_r. We adopt the "Cartesian" variant of (Seeger, 2009): candidates $\boldsymbol{X}_*$ are "phase encodes" (complete columns in Fourier space). MRI data has additional complexities (complex-valued $\boldsymbol{u}$, phase noise) which would interfere with our goals here: we employ a dataset of conventional natural images (the 75 images previously used in (Seeger & Nickisch, 2008)), at resolution 256 × 256. Measurement noise is fixed to $\sigma^2 = 10^{-3}$, the hyperparameters τ_a, τ_r adjusted based on MAP reconstructions with a fixed design picked adhoc ($\tau_a = 0.08, \tau_r = 0.16$). We split the data into five blocks, each containing one image for design optimization (Figure 1) and 5 test images drawn randomly from the pool. Sequential experimental design is run as previously done in (Seeger et al., 2009), both inference and MAP estimation use the same hyperparameters. Our test error measure is $\langle \|\boldsymbol{u}_* - \boldsymbol{u}_{\text{true}}\| / \|\hat{\boldsymbol{u}} - \boldsymbol{u}_{\text{true}}\| \rangle$, $\boldsymbol{u}_*$ the MAP estimate in question, $\hat{\boldsymbol{u}}$ the MAP image for a low-frequency-only design[4]. $\langle\cdot\rangle$ denotes test set and block averaging.

Figure 1. Training set for image acquisition optimization.

Recall K-Lanczos and L-PCA from Section 3. For inference with L-PCA, we can prove desirable properties (Section 4.2), yet it is not very practical for a number of reasons. First, there is no good reason for not using the full Lanczos representation after K steps: for L-PCA, we throw part of it away. Second, running Lanczos for K steps, we observe a significant fluctuation of the number of converged eigenvalues, certainly as long as designs are small. Finally, it is not obvious how to separate smallest and largest converged eigenvalues: the true spectrum of $\boldsymbol{A}$'s encountered tends to have gaps at both ends. We give L-PCA results here for comparison, but note that K-Lanczos will typically be used in practice.

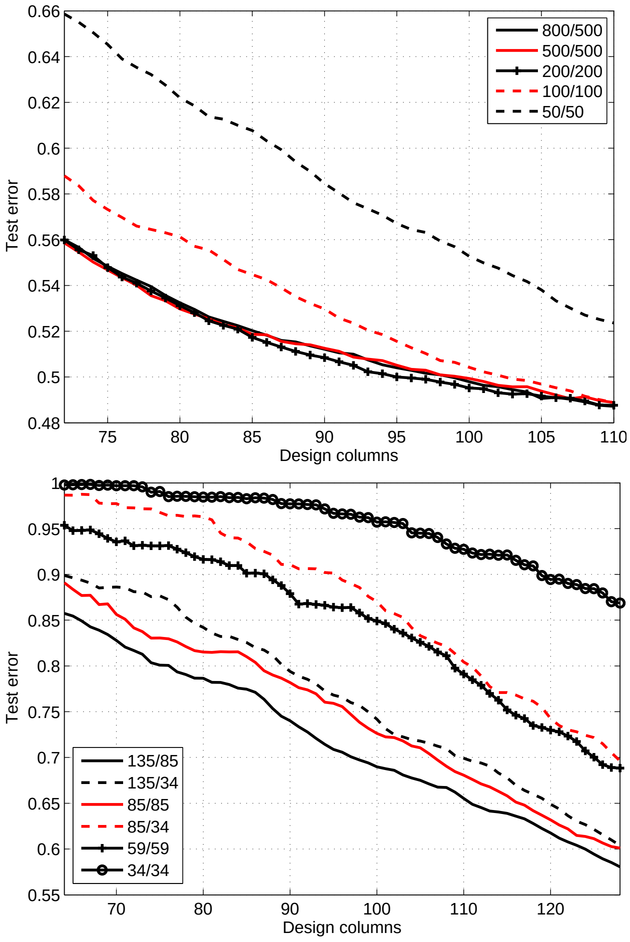

Figure 2. Relative test reconstruction errors for designs found with variational inference based on different-quality Gaussian variance approximations.
Top: K-Lanczos ($K_{\text{var}}/K_{\text{score}}$). Bottom: L-PCA ($L_{\text{var}}/L_{\text{score}}$).

Results are shown in Figure 2. First, K-Lanczos does much better than L-PCA[5]: the latter cannot be recommended. Results for K-Lanczos are remarkably robust across a wide range of Lanczos steps done. The curves between $K = 200$ and $K = 800$ differ insignificantly only (we show the most variable part). However, for too few Lanczos steps, results deteriorate (see also error images in Figure 3). Can we evaluate scores with less Lanczos iterations? We repeated the experiment with $K_{\text{var}} = \{500, 800\}$ (variances $\boldsymbol{\rho}$ at OL updates),

[3] Our goal is not compare the method of Seeger et al. (2009) against others, but to analyze effects of variance errors *for the same method*.

[4] The central 32 columns in Fourier space. Note that design optimization starts with this basis: it is contained in all other designs used here.

[5] Rather than trying to spot the "true central gap", we run Lanczos for
$K \approx 6L$ steps, obtaining $\tilde{N}$ converged eigenvalues, then use the $\min\{L, \tilde{N}\}$ smallest of these.

but $K_{\text{score}} = 200$, obtaining virtually the same results. For L-PCA, a similar maneuvre leads to further degradation (Figure 2, lower panel).

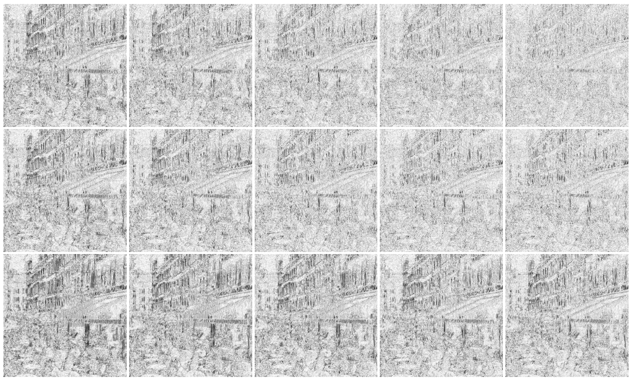

Figure 3. MAP reconstructions of training image under designs of 70, 80, 90, 100, 110 columns. Shown are residuals w.r.t. true image. Top: 200-Lanczos. Middle: 100-Lanczos. Bottom: 50-Lanczos.

These results underline our comments in Section 4.3. It should be noted that relative variance errors for K-Lanczos are of very different size between $K = 200$ and $K = 800$: the average relative error across all coefficients scales somewhat linearly in K (Seeger (2009), Sect. 5.1), a sizeable number of variance coefficients are orders of magnitude too small at $K = 200$. However, if posterior variances are used within a decision scenario, such as Bayesian experimental design for natural images, outcomes can be entirely robust in the presence of such errors.

6. Discussion

We have highlighted the significance of Gaussian variances approximation for variational (sparse) Bayesian inference and provided novel analyses about effects of PCA/Lanczos variance approximation errors on outcomes of nonlinear Bayesian experimental design. Our results show that outcomes can be robust in the presence of substantial overall marginal variance errors, at least for methods aiming to track dominating posterior covariances rather than imposing factorization constraints up front. While variational Bayesian methods are used in diverse applications, most evaluations of novel technology to date concentrate almost solely on closeness of marginals to the true posterior, a single point of merit which may often be of minor importance in practice. In order to understand real-world impact of Bayesian technology, theoretical analyses and empirical evaluations may have to broaden their focus.

ACKNOWLEDGEMENTS

This work is partly funded by the Excellence Initiative of the German research foundation (DFG). Thanks to Hannes Nickisch and David Wipf for helpful discussions.

References

Attias, H. A variational Bayesian framework for graphical models. In *NIPS 12*, pp. 209–215, 2000.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *JMLR*, 9:485–516, 2008.

Davis, C. All convex invariant functions of Hermitian matrices. *Archiv der Mathematik*, 8:276–278, 1957.

Golub, G. and Van Loan, C. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

Levin, A., Weiss, Y., Durand, F., and Freeman, W. Understanding and evaluating blind deconvolution algorithms. In *CVPR*, 2009.

Malioutov, D., Johnson, J., Choi, M., and Willsky, A. Low-rank variance estimation in GMRF models: Single and multiscale approaches. *IEEE Trans. Sig. Proc.*, 56(10):4621–4634, 2008.

Nickisch, H. and Seeger, M. Convex variational Bayesian inference for large scale generalized linear models. In *ICML 26*, pp. 761–768, 2009.

Schneider, M. and Willsky, A. Krylov subspace estimation. *SIAM J. Comp.*, 22(5):1840–1864, 2001.

Seeger, M. Bayesian inference and optimal design for the sparse linear model. *JMLR*, 9:759–813, 2008.

Seeger, M. Speeding up magnetic resonance image acquisition by Bayesian multi-slice adaptive compressed sensing. In *NIPS 22*, pp. 1633–1641, 2009.

Seeger, M. and Nickisch, H. Compressed sensing and Bayesian experimental design. In *ICML 25*, 2008.

Seeger, M., Nickisch, H., Pohmann, R., and Schölkopf, B. Bayesian experimental design of magnetic resonance imaging sequences. In *NIPS 21*, pp. 1441–1448, 2009.

Willsky, A. Multiresolution Markov models for signal and image processing. *Proc. IEEE*, 8(90):1396–1458, 2002.

Wipf, D. and Nagarajan, S. A new view of automatic relevance determination. In *NIPS 20*, 2008.