Analysis of VB Factorizations for Sparse and Low-Rank Estimation

David Wipf DAVIDWIPF@GMAIL.COM
Microsoft Research, Beijing

Abstract

Variational Bayesian (VB) factorial approximations anchor a wide variety of probabilistic models, where tractable posterior inference is almost never possible. This basic strategy is particularly attractive when estimating structured low-dimensional models of high-dimensional data, exemplified by the search for minimal rank and/or sparse approximations to observed data. To this end, VB models are frequently deployed across applications including multi-task learning, robust PCA, subspace clustering, matrix completion, affine rank minimization, source localization, compressive sensing, and assorted combinations thereof. Perhaps surprisingly however, there exists almost no attendant theoretical explanation for how various VB factorizations operate, and in which situations one may be preferable to another. We address this relative void by comparing arguably two of the most popular factorizations, one built upon Gaussian scale mixture priors, the other bilinear Gaussian priors, both of which can favor minimal rank or sparsity depending on the context. More specifically, by re-expressing the respective VB objective functions, we weigh multiple factors related to local minima avoidance, feature transformation invariance and correlation, and computational complexity to arrive at insightful conclusions useful in explaining performance and deciding which VB flavor is advantageous. We also envision that the principles explored here are quite relevant to other structured inverse problems where VB serves as a viable solution.

1. INTRODUCTION

Variational Bayesian (VB) techniques (Attias, 2000; Bishop, 2006) are now commonplace for performing approximate inference in structured probabilistic models of high-dimensional data, usually representing the only tractable alternative to computationally expensive MCMC sampling. Unfortunately however, even with a specific data set and operational goal, there is often no established prescription for how to optimally apply VB, and different parameterizations and associated distributional factorizations can have wide-ranging impacts. To elucidate these differences, we will consider observational models of the form

$$\boldsymbol{y} \;=\; \sum_{i=1}^{m} \mathcal{A}_i(X_i) + \boldsymbol{e} \;\equiv\; \sum_{i=1}^{m} \Phi_i \text{vec}[X_i] + \boldsymbol{e}, \quad (1)$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is an observed data vector, $\{X_i\}$ represents a set of unknown latent matrices we would like to estimate,[1] and $\boldsymbol{e}$ is noise with distribution $\mathcal{N}(\mathbf{0}, \lambda I)$, $\lambda > 0$. Additionally, $\{\mathcal{A}_i\}$ represents a set of known linear operators defined by the corresponding matrices $\{\Phi_i\}$, and vec$[\cdot]$ denotes the column-wise vectorization of a matrix. We also assume $X_i \in \mathbb{R}^{p_i \times \rho_i}$ for all i, and without loss of generality that $p_i \leq \rho_i$. Also, if $p_i = \rho_i = 1$ for some i, we may include scalars as a special case. Compounding the estimation difficulty is the allowance that the combined dimensionality of all $\{X_i\}$ may be substantially larger than $\boldsymbol{y}$, or $\sum_i p_i \rho_i \gg n$, meaning the measurement process is severely underdetermined or only partially observable. Consequently prior assumptions are required to restrict the solution space if we hope to recover $\{X_i\}$.

In this work we will be assuming that each unknown element X_i has relatively small rank, possibly even zero for many indeces, such that overall $\sum_i \text{rank}[X_i]$, or some weighted alternative, is minimal. While perhaps deceptively simple on the surface, special cases encompass a wide range of machine learning applications and data analysis tools including trace regression (Rohde & Tsybakov, 2011), multi-task learning (Jalali et al., 2010), robust PCA (Candès et al., 2011), subspace clustering (Liu et al., 2013), matrix completion (Candès & Recht, 2009), general affine rank minimization (Chandrasekaran et al., 2012), source localization (Limpiti et al., 2006), compressive sensing (Candès et al., 2006), and assorted combinations thereof (McCoy & Tropp, 2013; Nakajima et al., 2013a). Moreover, we envision that the underlying VB principles we

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

[1] We frequently adopt the abbreviated notation $\{X_i\}$ to denote the set $\{X_i : i \in \mathcal{I}\}$, with $\mathcal{I}$ an index set which should be apparent by context.

intend to dissect here will nonetheless remain relevant to other types of structured inverse problems.

Arguably the most direct approach to estimating the unknown set $\{X_i\}$ is to solve a regularized regression problem of the form

$$\min_{\{X_i\}} \|\boldsymbol{y} - \textstyle\sum_i \Phi_i \text{vec}[X_i]\|_2^2 + \sum_i \lambda_i \text{rank}[X_i], \quad (2)$$

where each $\lambda_i > 0$ is a weighting factor. For example, if X_i is some scalar x_i, Φ_i is an n-dimensional column vector $\boldsymbol{\phi}_i$, and $\lambda_i = \lambda$ for all i, then (2) reduces to the canonical sparse estimation problem

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \Phi \boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_0, \quad (3)$$

where $\Phi = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_m]$, $\boldsymbol{x} = [x_1, \ldots, x_m]^\top$, and $\|\boldsymbol{x}\|_0$ denotes the ℓ_0 norm, or a count of the number of nonzero elements in a vector. In contrast, if we instead assume a single matrix component $X \in \mathbb{R}^{p \times \rho}$ and design matrix $\Phi \in \mathbb{R}^{n \times p\rho}$, then (2) simplifies to

$$\min_{X} \|\boldsymbol{y} - \Phi \text{vec}[X]\|_2^2 + \lambda \text{rank}[X]. \quad (4)$$

If each row of Φ is all zeros with a lone '1,' then (4) is equivalent to the well-studied matrix completion problem (Candès & Recht, 2009); alternatively if elements of Φ are drawn iid from $\mathcal{N}(0,1)$ we have the Gaussian matrix recovery problem (Chandrasekaran et al., 2012).

While foundational to numerous application domains, all of these cases require an NP-hard optimization problem. This has of course prompted a wide variety of convex relaxations or other efficient approximations for practical deployment. But within this high-level context, and especially when narrowed to many special cases, variational Bayesian (VB) methods provide an attractive alternative that often outperform purely optimization-based methods.

Obviously (1) defines a likelihood function $p(\boldsymbol{y}|\{X_i\})$. From this starting point, VB algorithms can broadly be partitioned into two flavors based on how the underlying latent variables are defined and factored for inference purposes. The first is built upon Gaussian scale mixture representations (GSM) (Andrews & Mallows, 1974) of sparse or low-rank favoring prior distributions $p(X_i)$ (Bishop & Tipping, 2000; Neal, 1996; Palmer et al., 2006). More concretely, the prior is built as $\quad p(\{X_i\}) = \prod_i p(X_i), \quad p(X_i) =$

$$\int p(X_i|\Gamma_i)p(\Gamma_i)d\Gamma_i \propto \int \exp\left(-\tfrac{1}{2}\text{tr}\left[X_i^\top \Gamma_i X_i\right]\right) p(\Gamma_i)d\Gamma_i, \quad (5)$$

where $\Gamma_i \in \mathbb{R}^{p_i \times p_i}$ is a precision matrix with prior $p(\Gamma_i)$. In the scalar case, any prior of this form can be shown to be super-Gaussian (Palmer et al., 2006), with heavy tails and a sharp peak at zero (i.e., sparsity-promoting). In contrast, with suitable choice for $p(\Gamma_i)$ it can have the same effect on the singular values of X_i to favor minimal rank. Although we might like to compute the posterior $p(\{X_i\}|\boldsymbol{y})$, this is intractable; likewise for the expanded posterior $p(\{X_i\},\{\Gamma_i\}|\boldsymbol{y})$, where $\{\Gamma_i\}$ is the companion set of precisions defining the GSM prior for all X_i. Fortunately, VB offers a convenient means of computing surrogate posterior approximations.

The basic idea is to specify some class of distributions $q(\{X_i\},\{\Gamma_i\})$ and then minimize the Kullback-Leibler divergence $\text{KL}\left[q(\{X_i\},\{\Gamma_i\})||p(\{X_i\},\{\Gamma_i\}|\boldsymbol{y})\right]$ over $q(\{X_i\},\{\Gamma_i\})$ (Attias, 2000; Bishop, 2006). As it turns out, if $p(\Gamma_i)$ is a suitably chosen conjugate prior, in this case a Wishart distribution on each precision Γ_i, and we assume the factorization $q(\{X_i\},\{\Gamma_i\}) = q(\{X_i\})q(\{\Gamma_i\})$, then no additional restrictions need be applied and we can globally solve for $q(\{X_i\})$ with $q(\{\Gamma_i\})$ fixed and vice versa. Up to an irrelevant constant, this is equivalent to minimizing

$$\int q(\{X_i\})q(\{\Gamma_i\}) \log \frac{q(\{X_i\})q(\{\Gamma_i\})}{p\left(\boldsymbol{y},\{X_i\},\{\Gamma_i\}\right)} \textstyle\prod_i dX_i d\Gamma_i \quad (6)$$

over $q(\{X_i\})$ and $q(\{\Gamma_i\})$ sequentially via coordinate descent, a procedure we will refer to as *VB-GSM*. Upon convergence, the mean of $q(\{X_i\})$ is typically used as a final point estimate.

In contrast, a bilinear Gaussian (BG) factorization underscores the second type of VB model (Babacan et al., 2012; Ilin & Raiko, 2010; Lim & Teh, 2007; Nakajima et al., 2013a; Wang & Yeung, 2013). Here we generically assume that $X_i = A_i B_i^\top$ for matrices $A_i \in \mathbb{R}^{p_i \times r_i}$ and $B_i \in \mathbb{R}^{\rho_i \times r_i}$, and then adopt the independent Gaussian priors

$$\begin{aligned} p(A_i) &\propto \exp\left(-\tfrac{1}{2}\text{tr}\left[A_i \Omega_{A_i} A_i^\top\right]\right) \\ p(B_i) &\propto \exp\left(-\tfrac{1}{2}\text{tr}\left[B_i \Omega_{B_i} B_i^\top\right]\right), \end{aligned} \quad (7)$$

where Ω_{A_i} and Ω_{B_i} are diagonal precision matrices, possibly known (Mnih & Salakhutdinov, 2008; Nakajima et al., 2013b). In the special case where $\Omega_{A_i} = \Omega_{B_i} = \alpha_i I$, for some $\alpha_i > 0$, the corresponding induced prior on X_i can be shown to be $p(X_i) \propto \exp\left(-\alpha_i \|X_i\|_*\right)$ (Srebro et al., 2005), where $\|\cdot\|_*$ denotes the nuclear norm, a common convex penalty for rank minimization. Proceeding to VB inference, we now employ a factorization over the sets $\{A_i\}$ and $\{B_i\}$ and minimize

$$\int q(\{A_i\})q(\{B_i\}) \log \tfrac{q(\{A_i\})q(\{B_i\})}{p(\boldsymbol{y},\{A_i\},\{B_i\})} \textstyle\prod_i dA_i dB_i, \quad (8)$$

over $q(\{A_i\})$ and $q(\{B_i\})$, which we henceforth denote as *VB-BG*. If desired, this expression can also be optimized over precision sets $\{\Omega_{A_i}\}$ and $\{\Omega_{B_i}\}$ as well, a process referred to as empirical Bayes (Nakajima et al., 2013b).

Despite the widespread use of both VB-GSM and VB-BG templates, there exists essentially no existing *comparative* analysis of meaningful intrinsic differences, nor clear guidelines for which might lead to better performance. Additionally, VB-BG is especially devoid of supporting theoretical understanding except in one very particular special case, namely, when there is only a single unknown latent X and we observe all entries corrupted simply with iid Gaussian noise, or equivalently, the observation model (1) is simplified to

$$\boldsymbol{y} = \text{vec}[X] + \boldsymbol{e}. \tag{9}$$

In this restricted regime (see (Nakajima et al., 2013b) and many references within), often referred to as the *fully-observable* model since $n = p\rho$, the VB-BG estimator, like many other estimators Bayesian or otherwise, naturally reduces to $\hat{X} = Ug(S)V^\top$, where $USV^\top$ is the SVD of $\boldsymbol{y}$ (when stacked into a matrix in correspondence with X) and g is a shrinkage/thresholding operator on the respective singular values. Hence the behavior of VB-BG largely defaults to that of standard estimators, and has been very precisely characterized (Nakajima et al., 2013b). However, none of these results are suggestive of how VB-BG might perform once we move to the wider, practical class of NP-hard *partially-observable* models such as those encountered in sparse estimation, matrix recovery, robust PCA, and beyond that we intend to tackle here.

Our contributions are summarized as follows:

- In Section 2 we examine VB-GSM and VB-BG solutions to solving (3), where the number of components m is large but the dimensionality of each component, $p_i \times \rho_i = 1 \times 1$, is small (scalars). This analysis reveals a key differentiating factor that surfaces based on the structure of $\Phi^\top\Phi$, as well as connections with convex ℓ_1-norm penalized regression. Our results indicate that VB-GSM will be preferable when columns of Φ display some degree of correlation structure and maximal sparsity is paramount.

- In Section 3 we expose complementary theoretical differences on general matrix recovery of the form given by (4), where now $m = 1$ (small), but p and ρ are arbitrarily large. Here we show that tractable implementation of VB-BG requires an additional row-wise factorization, the net result being that the underlying cost function collapses to a standard regularized regression problem with many troublesome local minima, meaning that minimal rank solutions can be more evasive.

- In Section 4 we present corroborating empirical phenomena exactly predicted by our theory.

We emphasize that although our focus is on two illustrative cases, (I) sparse estimation and (II) affine matrix recovery, more complex models in the general from of (1) will naturally display attributes of these two building blocks. Hence any undesirable local minima properties or the sensitivity to design correlations will likely be inherited, extending the scope of our analysis. Moreover, if we choose a wider class of likelihood models (e.g., exponential families, etc.), then VB can no longer be applied in the present context without including an additional variational bound such as that used by (Jaakkola & Jordan, 2000). But once this bound is applied, then the objective function reduces to the same basic structures we already analyze herein, albeit with heteroscedastic noise, a relatively inconsequential difference.

2. CASE I: SPARSE ESTIMATION

This section will examine the basic sparse linear inverse problem from (3) from first the VB-GSM viewpoint and later that of VB-BG. In both cases the relevant likelihood function is

$$p(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\left[-\tfrac{1}{2\lambda}\|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2\right]. \tag{10}$$

In contrast, the assumed prior distribution will depend on whether we are applying VB-GSM or VB-BG leading to divergent algorithmic properties as discussed in Sections 2.1 and 2.2 respectively.

2.1. Gaussian Scale Mixture Factorization

The VB-GSM prior from (5) simplifies to $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})d\boldsymbol{\gamma} \propto \prod_i \int \exp\left(-\frac{\gamma_i x_i^2}{2}\right) p(\gamma_i)d\gamma_i$, where $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_m]^\top$ is a vector of unknown prior precisions. Each element is independently distributed as $p(\gamma_i) = \text{Gam}[c, d] \propto \gamma_i^{(c-1)} \exp[-d\gamma_i]$ for all i, or a conjugate Gamma distribution with non-negative parameters c and d fixed and known. The goal is then to optimize (6) over the factorized approximate distributions $q(\boldsymbol{x})$ and $q(\boldsymbol{\gamma})$, where now $\{X_i\} = \boldsymbol{x}$ and $\{\Gamma_i\} = \boldsymbol{\gamma}$. Obviously this objective involves functional-valued arguments and high-dimensional integrals. However, using standard VB results from (Bishop, 2006) we can actually convert to a more familiar parameterized form as follows. First, the optimal $q(\boldsymbol{x})$ and $q(\boldsymbol{\gamma})$ must satisfy

$$\begin{aligned} q(\boldsymbol{x}) &\propto \exp\left(\text{E}_{q(\boldsymbol{\gamma})}\left[\log p\left(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}\right)\right]\right) \quad \Rightarrow \quad \mathcal{N}(\bar{\boldsymbol{x}}, \Sigma_x), \\ q(\boldsymbol{\gamma}) &\propto \exp\left(\text{E}_{q(\boldsymbol{x})}\left[\log p\left(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}\right)\right]\right) \quad \Rightarrow \\ &\qquad \prod_i \text{Gam}\left(c + 0.5, \tfrac{c+0.5}{\bar{\gamma}_i}\right), \end{aligned} \tag{11}$$

which form the basis of the update rules for the popular variational relevance vector machine (Bishop & Tipping, 2000). Given these distributions, we observe that $q(\boldsymbol{x})$ only depends on a mean $\bar{\boldsymbol{x}}$ and a covariance Σ_x, while $q(\boldsymbol{\gamma})$ only depends on a mean parameter $\bar{\boldsymbol{\gamma}} \triangleq [\bar{\gamma}_1, \ldots, \bar{\gamma}_m]^\top$. Consequently, the VB-GSM problem for sparse estimation reduces to minimizing

$$\begin{aligned} \mathcal{L}(\bar{\boldsymbol{x}}, \Sigma_x, \bar{\boldsymbol{\gamma}}) &= \int q(\boldsymbol{x})q(\boldsymbol{\gamma}) \log \frac{q(\boldsymbol{x})q(\boldsymbol{\gamma})}{p\left(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}\right)} d\boldsymbol{x}d\boldsymbol{\gamma} \\ &= \text{E}_{q(\boldsymbol{x})q(\boldsymbol{\gamma})}\left[-\log p\left(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}\right)\right] - \text{H}\left[q(\boldsymbol{x})q(\boldsymbol{\gamma})\right] \end{aligned} \tag{12}$$

over $\bar{\boldsymbol{x}}$, Σ_x, and $\bar{\boldsymbol{\gamma}}$, where H denotes the entropy function.[2] Interestingly, with $\bar{\boldsymbol{\gamma}}$ fixed, this can be solved in closed-form, leading the following result.

Proposition 2.1. *Let* $\mathcal{L}(\bar{\boldsymbol{\gamma}}) \triangleq \inf_{\bar{\boldsymbol{x}},\Sigma_x} \mathcal{L}(\bar{\boldsymbol{x}}, \Sigma_x, \bar{\boldsymbol{\gamma}})$*. Then*

$$\begin{aligned} \mathcal{L}(\bar{\boldsymbol{\gamma}}) \;&=\; \boldsymbol{y}^\top \left(\lambda I + \Phi diag[\bar{\boldsymbol{\gamma}}]^{-1}\Phi^\top\right)^{-1} \boldsymbol{y} \\ &+\; \log\left|\lambda I + \Phi diag[\bar{\boldsymbol{\gamma}}]^{-1}\Phi^\top\right| + 2\textstyle\sum_i \left(d\bar{\gamma}_i - c\log\bar{\gamma}_i\right). \end{aligned} \tag{13}$$

Additionally, let $\bar{\boldsymbol{\gamma}}^*$ *be a minimum of* $\mathcal{L}(\bar{\boldsymbol{\gamma}})$ *and define*

$$\begin{aligned} \bar{\boldsymbol{x}}^* \;&\triangleq\; diag[\bar{\boldsymbol{\gamma}}^*]^{-1}\Phi^\top \left(\lambda I + \Phi diag[\bar{\boldsymbol{\gamma}}^*]^{-1}\Phi^\top\right)^{-1} \boldsymbol{y} \\ \Sigma_x^* \;&\triangleq\; \left(\lambda\Phi^\top\Phi + diag[\bar{\boldsymbol{\gamma}}^*]\right)^{-1}. \end{aligned} \tag{14}$$

Then $q(\boldsymbol{x}) = \mathcal{N}(\bar{\boldsymbol{x}}^*, \Sigma_x^*)$ *and* $q(\boldsymbol{\gamma}) = \prod_i Gam\left(c + 0.5, \frac{c+0.5}{\bar{\gamma}_i^*}\right)$ *collectively minimize (6).*

In the non-informative (or improper) limit of the hyperprior on $\boldsymbol{\gamma}$ where $c, d \to 0$, i.e., a flat prior in log space for scale parameters as motivated in (Tipping, 2001), the objective function $\mathcal{L}(\bar{\boldsymbol{\gamma}})$ can be derived differently from a type-II maximum likelihood perspective and has been analyzed extensively in the context of sparse estimation (Wipf et al., 2011; Aravkin et al., 2014); hence VB-GSM directly inherits all desirable attributes derived therein. To briefly summarize, minimizing $\mathcal{L}(\bar{\boldsymbol{\gamma}})$ maintains a number of intrinsic advantages over traditional sparse regression techniques such as ℓ_1-norm minimization or general MAP estimation. In particular, sparse solutions $\bar{\boldsymbol{x}}^*$ are guaranteed (as elements of $\bar{\boldsymbol{\gamma}}^*$ are pushed towards ∞), and in certain circumstances these solutions are guaranteed to be as sparse or sparser than the minimum ℓ_1 norm solution. Additionally, VB-GSM is invariant to a large degree of correlation structure among the columns of Φ, unlike most other existing estimators. Section 4 below will verify this empirically.

2.2. Bilinear Gaussian Factorization

For VB-BG we first assume that $\boldsymbol{x} = \boldsymbol{a} \odot \boldsymbol{b}$, where $\odot$ denotes the Hadamard product. The prior $p(\boldsymbol{x})$ is then specified by the priors $p(\boldsymbol{a}) \propto \exp\left[-\frac{1}{2}\boldsymbol{a}^\top\Omega_a\boldsymbol{a}\right]$ and $p(\boldsymbol{b}) \propto \exp\left[-\frac{1}{2}\boldsymbol{b}^\top\Omega_b\boldsymbol{b}\right]$, with Ω_a and Ω_b diagonal. Again, by adhering to standard VB procedures, we observe that the optimal approximate distributions will be of the form $q(\boldsymbol{a}) = \mathcal{N}(\bar{\boldsymbol{a}}, \Sigma_a)$ and $q(\boldsymbol{b}) = \mathcal{N}(\bar{\boldsymbol{b}}, \Sigma_b)$ for some respective means and covariances $\{\bar{\boldsymbol{a}}, \bar{\boldsymbol{b}}, \Sigma_a, \Sigma_b\}$. After applying these optimal parameterizations to the corresponding cost (8), performing a few algebraic manipulations, VB-BG can be shown to be equivalent to minimizing

$$\begin{aligned} \mathcal{L}(\bar{\boldsymbol{a}}, \bar{\boldsymbol{b}}, \Sigma_a, \Sigma_b) \;&\triangleq\; \tfrac{1}{\lambda}\|\boldsymbol{y} - \Phi(\bar{\boldsymbol{a}} \odot \bar{\boldsymbol{b}})\|_2^2 \\ &+\tfrac{1}{\lambda}\mathrm{tr}\left[\left(\Sigma_a \odot (\bar{\boldsymbol{b}}\bar{\boldsymbol{b}}^\top) + \Sigma_b \odot (\bar{\boldsymbol{a}}\bar{\boldsymbol{a}}^\top) + \Sigma_a \odot \Sigma_b\right)\Phi^\top\Phi\right] \\ &+\bar{\boldsymbol{a}}^\top\Omega_a\bar{\boldsymbol{a}} + \bar{\boldsymbol{b}}^\top\Omega_b\bar{\boldsymbol{b}} + \mathrm{tr}\left[\Omega_a\Sigma_a + \Omega_b\Sigma_b\right] - \log|\Sigma_a\Sigma_b| \end{aligned} \tag{15}$$

over $\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{b}}$, Σ_a, and Σ_b (this expression can be obtained by following a similar procedure as outlined in the proof of Proposition 2.1). At any minimizing solution, $\bar{\boldsymbol{a}}$ and $\bar{\boldsymbol{b}}$ are the posterior mean and $\bar{\boldsymbol{x}} \triangleq \bar{\boldsymbol{a}} \odot \bar{\boldsymbol{b}}$ acts as a final point estimate. Unlike in Proposition 2.1, there does not appear to be any direct way to collapse (15) to a convenient closed-form function of a single variable amenable to analysis. However, close inspection of targeted special cases reveals important underlying properties relevant to the sparse estimation problem and direct connections with VB-GSM.

Non-Informative Limit: We first consider the non-informative limit as $\Omega_a = \Omega_b = \alpha I$, with $\alpha \to 0$ (meaning a flat prior), reflecting maximum prior agnosticism regarding both $\boldsymbol{a}$ and $\boldsymbol{b}$ (Nakajima et al., 2013b). This can be viewed as the analogous non-informative limit of VB-GSM described above which has lead to considerable practical success.

Proposition 2.2. *Let* $\bar{\boldsymbol{x}} \triangleq \bar{\boldsymbol{a}} \odot \bar{\boldsymbol{b}}$ *such that* $\bar{\boldsymbol{a}} = \bar{\boldsymbol{x}}\bar{B}^{-1}$ *with* $\bar{B} \triangleq diag[\bar{\boldsymbol{b}}]$*. Also, let* $\mathcal{L}(\bar{\boldsymbol{b}}, \Sigma_b) \triangleq \inf_{\bar{\boldsymbol{x}},\Sigma_a} \lim_{\alpha\to\infty} \mathcal{L}(\bar{\boldsymbol{x}}\bar{B}^{-1}, \bar{\boldsymbol{b}}, \Sigma_a, \Sigma_b)$ *and* $W \triangleq \frac{1}{\lambda}\bar{B}^{-1}\Sigma_b\bar{B}^{-1}$*. Then* $\mathcal{L}(\bar{\boldsymbol{b}}, \Sigma_b) \equiv$

$$\begin{aligned} &\mathcal{L}(W) \triangleq \boldsymbol{y}^\top\left(\lambda I + \Phi\left(W \odot \Phi^\top\Phi\right)^{-1}\Phi^\top\right)\boldsymbol{y} \\ &+\log\left(\left|\lambda I + \Phi\left(W \odot \Phi^\top\Phi\right)^{-1}\Phi^\top\right|\left|W \odot \Phi^\top\Phi\right|/|W|\right). \end{aligned} \tag{16}$$

In words, this result implies that we can optimize (15) over both $\bar{\boldsymbol{x}}$ and Σ_a analytically, and once we have done so, the remaining dependency on $\bar{\boldsymbol{b}}$ and Σ_b can be merged into a single variable W. This simplification relates to the VB-GSM representation directly as follows:

Proposition 2.3. *If either (i)* W *is restricted to be diagonal, or (ii)* $\Phi^\top\Phi = I$*, then (16) is equivalent to the VB-GSM-based cost function from (13) in the non-informative limit* $c, d \to 0$*.*

Hence VB-BG will inherit all of the same sparsity-promoting properties as the VB-GSM representation mentioned in Section 2.1. Additionally, even without the assumption that $\Phi^\top\Phi = I$ (or that W is diagonal) it appears that under relatively mild conditions, and in the limit $\lambda \to 0$, both VB-BG and VB-GSM objective functions will have the same global solution (details omitted for brevity).

Crucially however, in general the overall objective function shape and local minima structure of VB-BG and VB-GSM

[2] For notational simplicity, we define all objective functions using $\mathcal{L}(\cdot)$, with different objectives being distinguished by the associated differing arguments and should be differentiable by context, similar to how $p(\cdot)$ is frequently used to define various distributions.

will be quite different as we transfer away from the simplified case with orthogonal designs, and it behooves us to probe further to better understand intrinsic differences that may influence sparse estimation. In particular, as correlations are introduced into Φ, the behavior of VB-BG and VB-GSM deviates sharply. The following result speaks to this effect:

Proposition 2.4. *Let* $\widetilde{\boldsymbol{y}} \triangleq \Psi \boldsymbol{y}$ *and* $\widetilde{\Phi} \triangleq \Psi\Phi$, *where* Ψ *is an arbitrary invertible matrix. Define the associated VB-BG objective function with* $\lambda \to 0$ *as*

$$\mathcal{L}(W;\Psi) \;\triangleq\; \widetilde{\boldsymbol{y}}^\top \left(\widetilde{\Phi} \left[W \odot \widetilde{\Phi}^\top \widetilde{\Phi} \right]^{-1} \widetilde{\Phi}^\top \right)^{-1} \widetilde{\boldsymbol{y}} \qquad (17)$$

$$+ \log \left| \widetilde{\Phi} \left[W \odot \widetilde{\Phi}^\top \widetilde{\Phi} \right]^{-1} \widetilde{\Phi}^\top \right| + \log \left| W \odot \widetilde{\Phi}^\top \widetilde{\Phi} \right| - \log |W|.$$

Unlike the corresponding VB-GSM cost from (13), $\mathcal{L}(W;\Psi)$ is *not* invariant *to transformations via* Ψ. *As an extreme case, in the limit as* Ψ *approaches any rank one matrix* $\boldsymbol{u}\boldsymbol{v}^\top$, *we have that*

$$\lim_{\Psi \to \boldsymbol{u}\boldsymbol{v}^\top} \mathcal{L}(W;\Psi) \equiv \boldsymbol{y}^\top \left(\Phi W^{-1} \Phi^\top \right)^{-1} \boldsymbol{y} + \log \left| \Phi W^{-1} \Phi^\top \right| \qquad (18)$$

up to irrelevant constant scaling and additive factors.

Proposition 2.4 elucidates an important distinction between VB-BG and VB-GSM. First, as $\Phi^\top\Phi$ transitions from an identity matrix (with no correlation structure) to a rank one matrix, the VB-BG cost actually becomes *less* encouraging of sparse solutions, and in the limiting case, sparsity is *not favored at all*. To see this, observe that (18) can be driven to $-\infty$ using *any* $W = USU^\top$ such that $\boldsymbol{y} \in \text{span}[\Phi U S^{-1/2}]$ and S^{-1} with (approximately) rank less than n. Under these conditions, the data term can be held constant while the log-det can be driven to an arbitrarily large negative value. But this will not typically lead to a sparse estimate of $\boldsymbol{x}$. In general, given any optimal W^*, the associated optimal $\bar{\boldsymbol{x}}^*$ is computed via (14), with $\left(W^* \odot \Phi^\top\Phi\right)$ replacing $\text{diag}[\bar{\gamma}^*]$. The sparsity of VB-GSM comes from the sparse diagonal structure of this $\text{diag}[\bar{\gamma}^*]^{-1}$; however, with VB-BG $\left(W^* \odot \Phi^\top\Phi\right)^{-1}$ need not be either sparse nor diagonal. In some ways this behavior is similar to the trace Lasso estimator, which behaves like the minimum ℓ_1-norm solution with an orthogonal design matrix, and the minimum ℓ_2-norm solution with a (nearly) rank one design (Grave et al., 2011).

Additionally, there is nothing intrinsically special about introducing correlations into Φ via left multiplication by Ψ. In general, the more $\Phi^\top\Phi$ contains decaying singular values, the more $W \odot \Phi^\top\Phi$ behaves like W, and the lesser the preference for sparse solutions exhibited by VB-BG. Simulation experiments in Section 4 below explicitly illustrate this effect. Finally, although not explored here for space considerations, when we learn Ω_a and Ω_b via empirical Bayes, the resulting cost function has similar behavior (both theoretically and empirically) to the non-informative limiting case, and therefore does not represent a solution to the drawbacks of VB-BG discussed in this section.

Strongly Informative Limit: Conversely, we now consider $\Omega_a = \Omega_b = \alpha I$, with α becoming large. In general, the larger α, the stronger the prior forcing elements of $\boldsymbol{a}$ and $\boldsymbol{b}$, and therefore $\boldsymbol{x}$, towards zero. However, provided λ is reduced at a proportional rate, then the data fit term remains significant such that the degenerate solution $\boldsymbol{a} = \boldsymbol{b} = \boldsymbol{0}$ can be avoided and we can still isolate the effects of the prior. With these considerations in mind, we have the following:

Proposition 2.5. *In the limit* $\lambda \to 0, \alpha \to \infty$, *with* $\lambda\alpha \to C$ *for some constant* $C > 0$, *we have that*

$$\inf_{\Sigma_a, \bar{\boldsymbol{b}}} \mathcal{L}(\bar{\boldsymbol{a}}, \bar{\boldsymbol{b}}, \Sigma_a, \Sigma_b) \to \|\boldsymbol{y} - \Phi\bar{\boldsymbol{x}}\|_2^2 + 2C\|\bar{\boldsymbol{x}}\|_1, \qquad (19)$$

where $\bar{\boldsymbol{x}} \triangleq \bar{\boldsymbol{a}} \odot \bar{\boldsymbol{b}}$.

In brief, this implies that the VB-BG objective will lose its dependency on Σ_b and converge to standard ℓ_1-norm penalized regression given the stated limiting conditions on α. The experiments presented in Section 4 provide corroborating empirical evidence of this general trend.

3. CASE II: MATRIX RECOVERY FROM AFFINE MEASUREMENTS

In Section 2 we analyzed the situation where the number of components m was large, but the dimensionality of each component X_i was small (a scalar), and hence low rank equated with sparsity. In this section we turn things around and consider the complementary case where $m = 1$ (a single component), but the unknown $X \in \mathbb{R}^{p \times \rho}$ can have arbitrarily large dimensions. For both VB-GSM and VB-BG the appropriate likelihood function is given by

$$p(\boldsymbol{y}|X) \propto \exp\left[-\tfrac{1}{2\lambda}\|\boldsymbol{y} - \Phi\text{vec}[X]\|_2^2\right]. \qquad (20)$$

Although the likelihood function is more or less identical to the one described in Section 2, the prior distributions are quite different, with distinctive attendant analyses and algorithms, especially in the case of VB-BG.

3.1. Gaussian Scale Mixture Factorization

For VB-GSM, the prior is expressed as in (5), but with no subscript required as there is a single component. Also, $p(\Gamma) = \mathcal{W}_p(\Lambda, \beta) \propto |\Gamma|^{(\beta-p-1)/2}\exp[-\frac{1}{2}\text{tr}(\Lambda\Gamma)]$, a Wishart distribution with degrees-of-freedom parameter $\beta > p - 1$ and positive definite, symmetric scaling matrix Λ. We then follow a similar analysis pipeline as in Section 2.1, arriving at the optimal factorized distributions

$$q(\text{vec}[X]) = \mathcal{N}(\text{vec}[\bar{X}], I \otimes \Sigma_X), \qquad (21)$$
$$q(\Gamma) = \mathcal{W}_p\left([\rho + \beta]\,\bar{\Gamma}^{-1}, \rho + \beta\right),$$

where $\otimes$ denotes the Kronecker product, and the revised implicit cost function

$$\mathcal{L}(\bar{\Gamma}) \quad = \quad \boldsymbol{y}^\top \left(\lambda I + \Phi(I \otimes \bar{\Gamma})^{-1}\Phi^\top\right)^{-1} \boldsymbol{y} \; + \tag{22}$$
$$\log \left|\lambda I + \Phi(I \otimes \bar{\Gamma})^{-1}\Phi^\top\right| + \text{tr}\left[\Lambda\bar{\Gamma}\right] - (\beta - p + 1)\log|\bar{\Gamma}|$$

after analogous simplifications (details are somewhat redundant and omitted for brevity). This objective function is quite similar to (13) and in the non-informative limit as $(\beta - p + 1)$ and Λ go to zero, it has been analyzed to some extent (Xin & Wipf, 2015), again from the vantage point of type-II maximum likelihood. In general, it acts as a close proxy for the matrix rank function (maintaining the same global solution and scale invariance), but with many fewer bad locally minimizing solutions. It therefore represents a robust tool for solving rank minimization problems.

3.2. Bilinear Gaussian Factorization

VB-BG, which is more commonly applied to rank minimization and matrix recovery problems than VB-GSM, again follows a similar path as before. The relevant posteriors are

$$\begin{aligned} q(\text{vec}[A^\top]) &= \mathcal{N}\left(\text{vec}[\bar{A}^\top], \Sigma_A\right) \\ q(\text{vec}[B^\top]) &= \mathcal{N}\left(\text{vec}[\bar{B}^\top], \Sigma_B\right). \end{aligned} \tag{23}$$

The corresponding parameterized objective function analogous to (15) can be shown to be $\mathcal{L}(\bar{A}, \bar{B}, \Sigma_A, \Sigma_B) \triangleq$

$$\begin{aligned} &\tfrac{1}{\lambda}\|\boldsymbol{y} - \Phi\text{vec}\left[AB^\top\right]\|_2^2 + \tfrac{1}{\lambda}\text{tr}\Big[\Sigma_A\left(I \otimes \bar{B}\right)^\top \Phi^\top\Phi\left(I \otimes \bar{B}\right) \\ &+ \;\Sigma_B\left(I \otimes \bar{A}\right)^\top P\Phi^\top\Phi P\left(I \otimes \bar{A}\right)\Big] + \tfrac{1}{\lambda}f(\Sigma_A, \Sigma_B) \\ &+ \;\text{tr}\left[A\Omega_A A^\top + B\Omega_B B^\top\right] - \log|\Sigma_A\Sigma_B| \\ &+ \;\text{tr}\left[\Sigma_A\left(I \otimes \Omega_A\right) + \Sigma_B\left(I \otimes \Omega_B\right)\right], \end{aligned} \tag{24}$$

where f is a bilinear function of its arguments and P is a permutation matrix. For any reasonable degree of scalability, it is not feasible to work with full covariances for Σ_A and Σ_B, which will be $(pr) \times (pr)$ and $(\rho r) \times (\rho r)$ matrices respectively, and without knowledge of the true rank we must assume $r = \min(p, \rho)$. Hence the required VB algorithm would require inverting huge matrices.

Therefore, we instead assume that the approximate posterior distributions $q(A)$ and $q(B)$ factorize over rows, a standard assumption which is equivalent to requiring that Σ_A and Σ_B are block diagonal, with $r \times r$ dimensional blocks denoted $\{\Sigma_{A_i}\}$ and $\{\Sigma_{B_j}\}$ respectively. This factorization can be justified, at least in part, by the fact that when $\Phi^\top\Phi$ is diagonal (which includes the case of matrix completion), then the posterior automatically possesses these row-wise factorizations anyway.[3] Given this necessary assumption, (24) simplifies to

$$\begin{aligned} \mathcal{L}(\bar{A}, \bar{B}, \{\Sigma_{A_i}\}, \{\Sigma_{B_j}\}) \;&\triangleq\; \tfrac{1}{\lambda}\|\boldsymbol{y} - \Phi\text{vec}\left[AB^\top\right]\|_2^2 \\ &+ \tfrac{1}{\lambda}\textstyle\sum_i \text{tr}\left[\Sigma_{A_i}\bar{B}^\top G_i \bar{B}\right] + \tfrac{1}{\lambda}\textstyle\sum_j \text{tr}\left[\Sigma_{B_j}\bar{A}^\top H_j \bar{A}\right] \\ &+ \tfrac{1}{\lambda}f(\{\Sigma_{A_i}\}, \{\Sigma_{B_j}\}) + \text{tr}\left[A\Omega_A A^\top + B\Omega_B B^\top\right] \\ &+ \textstyle\sum_i \left(\text{tr}\left[\Sigma_{A_i}\Omega_A\right] - \log|\Sigma_{A_i}|\right) \\ &+ \textstyle\sum_j \left(\text{tr}\left[\Sigma_{B_j}\Omega_B\right] - \log|\Sigma_{B_j}|\right), \end{aligned} \tag{25}$$

where f is redefined accordingly (the exact form of f is not important for what follows, beyond its bilinearity), and $\{G_i\}$ and $\{H_j\}$ are the required block diagonal elements of $\Phi^\top\Phi$ and $P\Phi^\top\Phi P$ respectively. As before, we now consider relevant special cases more amenable to analysis.

Non-Informative Limit: Here we again assume that $\Omega_A = \Omega_B = \alpha I$, with $\alpha \to 0$, allowing us to ignore the α-dependent terms originating from the prior. But this pruned objective still remains difficult to unpack largely because of the term $f(\{\Sigma_{A_i}\}, \{\Sigma_{B_j}\})$ that couples the covariances in a bilinear yet complex way. However, if we also consider the limit as λ becomes small (canonical noiseless case), insightful analysis is possible based on the following result. We also assume A and B are square matrices for simplicity (i.e., $p = \rho$), although this can be relaxed with additional effort. We first define the indicator function $\mathcal{I}_\infty[z \neq \theta]$, which equals zero if $z = \theta$ or ∞ otherwise. Then we have the following:

Proposition 3.1. *Assume that $\alpha \to 0$, G_i and H_j are full rank for all i and j, and that $p = \rho$. Then excluding irrelevant constants/scale factors, at all full-rank $\bar{A}$ and $\bar{B}$*

$$\begin{aligned} &\lim_{\lambda \to 0} \inf_{\{\Sigma_{A_i}\}, \{\Sigma_{B_j}\}} \mathcal{L}(\bar{A}, \bar{B}, \{\Sigma_{A_i}\}, \{\Sigma_{B_j}\}) \\ &\equiv \log|\bar{A}\bar{A}^\top| + \log|\bar{B}\bar{B}^\top| + \mathcal{I}_\infty\left(\boldsymbol{y} \neq \Phi vec\left[\bar{A}\bar{B}^\top\right]\right). \end{aligned} \tag{26}$$

Technically speaking, Proposition 3.1 only applies when $\bar{A}$, $\bar{B}$, and therefore $\bar{X} \triangleq \bar{A}\bar{B}^\top$ are all full rank. However, except for perhaps infinitesimally small regions around low-rank solutions (and the set of low-rank matrices have Lebesgue measure zero), which are essentially irrelevant to the basic trajectory of the VB update rules anyway, the VB-BG is equivalent to solving

$$\min_{\bar{X}} \; \log|\bar{X}\bar{X}^\top| \quad \text{s.t. } \boldsymbol{y} = \Phi\text{vec}\left[\bar{X}\right]. \tag{27}$$

This only requires that we convert the $\mathcal{I}_\infty$ term to an equivalent constraint and apply $\log|\bar{X}\bar{X}^\top| = \log|\bar{A}\bar{A}^\top| + \log|\bar{B}\bar{B}^\top|$ to the objective. In fact VB updates are guaranteed to reduce or leave unchanged (27). At one level this is a reasonable endeavor given that $\log|\bar{X}\bar{X}^\top| \propto \sum_i \log \sigma_i[\bar{X}]$, where $\sigma_i[\bar{X}]$ denotes the i-th singular value of $\bar{X}$. Moreover, because $\log z = \lim_{\epsilon \to 0} \frac{1}{\epsilon}(z^\epsilon - 1)$, then $\sum_i \log \sigma_i[\bar{X}]$ approaches the rank function (up to constant scaling and translation), and can be viewed as a smooth proxy for favoring (nearly) low-rank solutions.

[3] The alternative column-wise factorization has other issues.

The problem however is that this penalty has a combinatorial number of locally minimizing solutions. This occurs because whenever any $\sigma_i[\bar{X}]$ approaches zero, the cost approaches $-\infty$ and we are necessarily trapped in a basin of attraction whereby this singular value can never become large again. In fact, minimization of (27) is NP-hard, and various smoothing heuristics to incrementally modify the cost function have already been suggested to improve performance (Mohan & Fazel, 2012). Unfortunately though, algorithms like VB-BG that directly minimize this function can be easily trapped, which likely explains modest performance relative to the algorithm from (Mohan & Fazel, 2012). Additionally, even when $p \neq \rho$, because the implicity penalty function only depends on G_i and H_j, not the entire Φ matrix, it will not be scale invariant either in the sense described in Section 2. And finally, analogous to the sparse estimation case, if we attempt to learn Ω_A and Ω_B, the implicit cost function of VB-BG is not significantly altered and hence the same problems remain.

Strongly Informative Limit: In the opposite extreme as α becomes large, VB-BG behavior mimics our findings for the sparse estimation case from Section 2.2. Under analogous limiting conditions, VB-BG converges to minimizing

$$\mathcal{L}(\bar{X}) \triangleq \tfrac{1}{\lambda}\|\boldsymbol{y} - \Phi \text{vec}\left[\bar{X}\right]\|_2^2 + 2C\|\bar{X}\|_*. \qquad (28)$$

We omit the full derivation, but the basic structure mirrors that of Proposition 2.5.

4. EMPIRICAL EXAMPLES

As our focus thus far has been on evaluating existing VB paradigms rather than developing new algorithms, the purpose of this section is restricted to merely presenting a few tailored simulations that complement the analytical conclusions described previously. To begin, we will explore some of the claims made in Section 2, which compares various flavors of VB sparse estimation algorithms. Specifically, we have revealed analytically that VB-BG in the non-informative limit is likely to be quite sensitive to correlations in the design matrix Φ, while VB-GSM (also in an analogous non-informative limit) is likely to be free of such concerns, consistent with existing work from (Wipf et al., 2011). Additionally, we demonstrated that VB-BG with a highly informative prior converges to something like an ℓ_1 norm regression problem, and hence we may expect its behavior to mirror the popular Lasso estimator in this regime. Finally, although not our primary emphasis, we mentioned that VB-BG, with a learned prior via empirical Bayes (Nakajima et al., 2013b) will likely exhibit some of the same shortcomings as the non-informative limit.

To illustrate these effects, we conducted the following Monte Carlo experiment. First we generate a sparse vector $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_0 = 20$ nonzero elements randomly located with iid $\mathcal{N}(0,1)$ nonzero elements. Next we generate a design matrix via $\Phi = \sum_{i=1}^{n} \frac{1}{i^\eta}\boldsymbol{u}_i\boldsymbol{v}_i^\top$, where each vector $\boldsymbol{u}_i \in \mathbb{R}^{50}$ and $\boldsymbol{v}_i \in \mathbb{R}^{100}$ are distributed iid with $\mathcal{N}(0,1)$ elements. We then normalized columns of Φ to have unit ℓ_2 norm. The exponent parameter η is chosen from the interval $[0,2]$, the effect being that larger values of η will introduce larger correlations into the resulting columns of Φ, meaning that $\Phi^\top\Phi$ will have stronger off-diagonal elements. Finally we generate a data vector via $\boldsymbol{y} = \Phi\boldsymbol{x}$. We then run various VB sparse estimation algorithms and evaluate them using two metrics, *normalized MSE* $\triangleq \left\langle \frac{\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2^2}{\|\boldsymbol{x}\|_2^2} \right\rangle$ and *average # nonzeros* $\triangleq \langle\|\hat{\boldsymbol{x}}\|_0\rangle$, where the empirical average $\langle\cdot\rangle$ is taken across 1000 independent trials. This process is repeated for values of $\eta \in [0,2]$, with results reported in Figure 1.

We observe that VB-BG in the non-informative limit (we chose $\alpha = 10^{-4}$) performs exactly as predicted by Proposition 2.4 and the attendant discussion that follows. When η is near zero, $\frac{1}{i^\eta} \approx 1$ for all i, and so $\Phi^\top\Phi$ has little correlation structure (although some small amount will exist due to natural statistical variability). Therefore, the performance of VB-GSM and VB-BG with their respective non-informative priors have nearly the same normalized MSE and sparsity level, and both fare far better than the standard Lasso estimator, which finds the minimum ℓ_1 norm solution. In contrast, when η becomes large and stronger correlations are introduced (as is common in many practical applications of sparse estimation), VB-BG performance degrades dramatically as anticipated, becoming much worse than even the Lasso estimator which is already well-know to be sensitive to such correlations (Candès et al., 2006). In fact, VB-BG (with non-informative prior) barely produces sparse estimates at all as seen in Figure 1 (*bottom*).

Note that by construction, with probability one we can obtain a feasible solution to $\boldsymbol{y} = \Phi\hat{\boldsymbol{x}}$ with $\|\hat{\boldsymbol{x}}\|_0 = 50$ using *any* subset of 50 columns of Φ (the maximum required). So once the average sparsity level approaches 50, a given algorithm is not accomplishing anything particularly challenging. Of course with $\eta > 0.6$, VB-BG with the non-informative prior exceeds even this minimal number of nonzeros, with $\|\hat{\boldsymbol{x}}\|_0$ approaching 100, or *a completely non-sparse solution* as predicted by our theory. Consequently, VB-BG is only advisable when maximal sparsity is not the prevailing concern, and it may be more applicable as a Bayesian competitor to the trace Lasso (Grave et al., 2011). In contrast, at the other extreme, when α becomes large for VB-BG (strong or informative prior case), the behavior approaches that of the Lasso, both in terms of the MSE recovery error and average sparsity, consistent with Proposition 2.5. Finally, although not our primary emphasis, we also include results for the empirical Bayesian version of VB-BG, where the prior parameters Ω_a and Ω_b are learned from the data. Performance is consistently worse than the Lasso, and the average value of $\|\hat{\boldsymbol{x}}\|_0$ can even exceed 50 as predicted. In fact, we can also explain why the performance of VB-BG with a learned prior deteriorates more slowly than VB-BG with a fixed non-informative prior, but this requires rather lengthy technical arguments which we prefer to leave for a future publication.

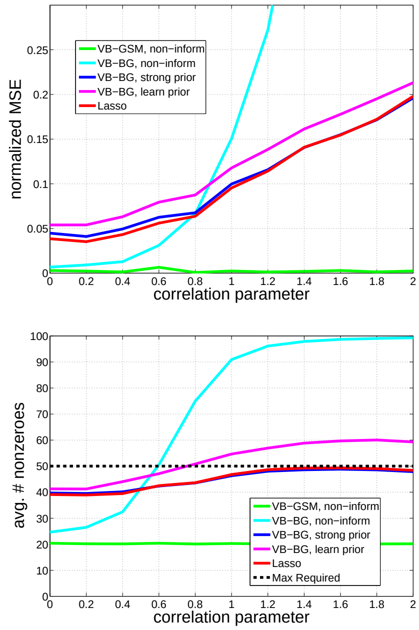

Figure 1. Top: Normalized MSE as correlation parameter η is varied. *Bottom*: Average number of nonzeros in $\hat{\boldsymbol{x}}$.

Turning to the matrix recovery problem from Section 3, we already know based on results in (Mohan & Fazel, 2012) that the implicit VB-BG penalty function $\log|XX^\top|$ derived in Section 3.2 may not be advisable for matrix recovery and a heuristic smoothing term is required to obtain more robust results. Additionally, although not labeled with this terminology, (Xin & Wipf, 2015) has already shown empirical comparisons between VB-BG, the smoothed version of $\log|XX^\top|$ from (Mohan & Fazel, 2012), and VB-GSM. In this context, VB-GSM performs the best while VB-BG lags the others by a significant margin as expected based on our theoretical results.

5. DISCUSSION

Often VB inference pipelines are mistakenly assumed to be more or less the same because they are based on similar structure-inducing priors and factorial approximations to the KL divergence from the true posterior. However, despite widespread practical use, there currently exists almost no rigorous analysis of the extent to which this is actually the case, or rather the degree to which we might prefer one flavor of VB to another in certain situations. In this work we have attempted to at least partially fill this void by examining two important cases that often form the building blocks of larger systems, revealing that VB-BG is likely to be more sensitive to design correlations and local minima, which can be problematic if maximally parsimonious representations are desired. However, on the positive side, by manually specifying a maximum rank r_i VB-BG does benefit from a direct route for injecting *a prior* information regarding the true rank of a given component (a flexibility not enjoyed by VB-GSM), and this can be exploited to increase computational efficiency.

Regarding broader generalizations, we have observed that the VB-GSM objective directly defaults to a relatively simple function based on a data-dependent inverse term and a log-det term. This then suggests that VB-GSM inherits desirable theoretical attributes previously ascribed to type-II maximum likelihood estimators in the context of sparse estimation, affine rank minimization, and robust PCA (Aravkin et al., 2014; Wipf, 2012; Wipf et al., 2011; Xin & Wipf, 2015). In fact, with some additional effort it can be shown that this simplification naturally generalizes to any likelihood function derived from (1) in the non-informative limit leading to the broadly-applicable VB-GSM objective

$$\mathcal{L}(\{\bar{\Gamma}_i\}) = \boldsymbol{y}^\top \left(\lambda I + \textstyle\sum_i \Phi_i (I \otimes \bar{\Gamma}_i)^{-1} \Phi_i^\top\right)^{-1} \boldsymbol{y} + \log\left|\lambda I + \textstyle\sum_i \Phi_i (I \otimes \bar{\Gamma}_i)^{-1} \Phi_i^\top\right|. \qquad (29)$$

When viewed from this vantage point, we may then expect to see expanded analytical studies of VB-GSM into other applications of interest that leverage the same model structure. Additionally, while not our primary focus, with judicious implementation we can optimize (29) with VB updates of worst-case order $O(n^2 \sum_i p_i \rho_i)$, or linear in the combined dimensionality of $\{X_i\}$ and quadratic in the number of observations.

In contrast, VB-BG admits no such general representation, and multiple potential drawbacks emerge from our analysis. One issue is that without additional factorizations, VB-BG typically faces a wider parameter space to search that includes problematic locales and/or local minima. For example, in the case of sparse estimation, we observed in Section 2.2 that the primary difference between VB-BG and VB-GSM was that the former had a full parameter matrix to estimate while the latter only required a diagonal one (i.e., vector). Furthermore, these extra degrees of freedom in off-diagonal terms granted VB-BG the undesirable ability to explore and actually favor non-sparse solutions that counter the original goal of finding maximally sparse estimates. A second issue is that, for computational reasons discussed in Section 3.2, VB-BG is likely to require additional heuristic factorizations of the approximate posterior $q(\{A_i\},\{B_i\})$, and such factorizations will increase the set of distracting candidate local minima. Interestingly, this observation provides further concrete confirmation of previous conjectures of this sort (Hoffman, 2014).

References

Andrews, D. F. and Mallows, C. L. Scale mixtures of normal distributions. *J. Royal Statistical Society Ser. B*, 36: 99–102, 1974.

Aravkin, A., Burke, J.V., Chiuso, A., and Pillonetto, G. Convex vs non-convex estimators for regression and sparse estimation: The mean squared error properties of ARD and GLasso. *J. Machine Learning Research*, 15: 217–252, 2014.

Attias, H. A variational Bayesian framework for graphical models. *Advances in Neural Information Processing Systems 12*, 2000.

Babacan, S.D., Luessi, M., Molina, R., and Katsaggelos, A.K. Sparse Bayesian methods for low-rank matrix estimation. *IEEE Trans. Signal Processing*, 60(8):3964–3977, 2012.

Bishop, C.M. *Pattern recognition and machine learning*. Springer, New York, 2006.

Bishop, C.M. and Tipping, M.E. Variational relevance vector machines. *Uncertainty in Artificial Intelligence*, 2000.

Candès, E. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Candès, E., Romberg, J., and Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Information Theory*, 52(2):489–509, 2006.

Candès, E., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *J. ACM*, 58(2), 2011.

Chandrasekaran, V., Recht, B., Parrilo, P., and Willsky, A. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

Grave, E., Obozinski, G.R., and Bach, F.R. Trace Lasso: A trace norm regularization for correlated designs. *Advances in Neural Information Processing Systems 24*, 2011.

Hoffman, M. Why variational inference gives bad parameter estimates. *Advances in Variational Inference, NIPS 2014 Workshop*, 2014.

Ilin, A. and Raiko, T. Practical approaches to principal component analysis in the presence of missing values. *J. Machine Learning Research*, 11:1957–2000, 2010.

Jaakkola, T. and Jordan, M. Bayesian parameter estimation through variational methods. *Statistics and Computing*, 10(1), 2000.

Jalali, A., Sanghavi, S., Ruan, C., and Ravikumar, P.K. A dirty model for multi-task learning. *Advances in Neural Information Processing Systems 23*, 2010.

Lim, Y.J. and Teh, Y.W. Variational Bayesian approach to movie rating prediction. *KDD Cup*, 2007.

Limpiti, T., Veen, B.D. Van, and Wakai, R.T. Cortical patch basis model for spatially extended neural activity. *IEEE Trans. Biomedical Engineering*, 53(9):1740–1754, 2006.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

McCoy, M.B. and Tropp, J.A. The achievable performance of convex demixing. *arXiv:1309.7478*, 2013.

Mnih, A. and Salakhutdinov, R.R. Probabilistic matrix factorization. *Advances in Neural Information Processing Systems 20*, 2008.

Mohan, K. and Fazel, M. Iterative reweighted algorithms for matrix rank minimization. *J. Machine Learning Research*, 13(1):3441–3473, 2012.

Nakajima, S., Sugiyama, M., and Babacan, S.D. Variational Bayesian sparse additive matrix factorization. *Machine Learning*, 92, 2013a.

Nakajima, S., Sugiyama, M., Babacan, S.D., and Tomioka, R. Global analytic solution of fully-observed variational Bayesian matrix factorization. *J. Machine Learning Research*, 14:1–37, 2013b.

Neal, R.M. *Bayesian Learning for Neural Networks*. Springer-Verlag, New York, 1996.

Palmer, J.A., Wipf, D.P., Kreutz-Delgado, K., and Rao, B.D. Variational EM algorithms for non-Gaussian latent variable models. *Advances in Neural Information Processing Systems 18*, 2006.

Rohde, A. and Tsybakov, A.B. Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 39 (2):887–930, 2011.

Srebro, Nathan, Rennie, Jason, and Jaakkola, Tommi S. Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems 17*, 2005.

Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

Wang, N. and Yeung, D.Y. Bayesian robust matrix factorization for image and video processing. *International Conference on Computer Vision*, 2013.

Wipf, D.P. Non-convex rank minimization via an empirical bayesian approach. *Uncertainty in Artificial Intelligence*, 2012.

Wipf, D.P., Rao, B.D., and Nagarajan, S. Latent variable Bayesian models for promoting sparsity. *IEEE Trans. Information Theory*, 57(9), 2011.

Xin, B. and Wipf, D.P. Pushing the limits of affine rank minimization by adapting probabilistic pca. *International Conference on Machine Learning*, 2015.