Variational Gaussian approximation for Poisson data

OPEN ACCESS
IOP Publishing
Inverse Problems

Inverse Problems **34** (2018) 025005 (29pp) https://doi.org/10.1088/1361-6420/aaa0ab

Variational Gaussian approximation for Poisson data

Simon R Arridge[1], **Kazufumi Ito**[2], **Bangti Jin**[1] and **Chen Zhang**[1]

[1] Department of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom
[2] Department of Mathematics, North Carolina State University, Raleigh, NC 27607, United States of America

E-mail: s.r.arridge@ucl.ac.uk, kito@ncsu.edu, b.jin@ucl.ac.uk and chen.zhang.16@ucl.ac.uk

Received 7 June 2017, revised 1 October 2017
Accepted for publication 11 December 2017
Published 12 January 2018

Abstract

The Poisson model is frequently employed to describe count data, but in a Bayesian context it leads to an analytically intractable posterior probability distribution. In this work, we analyze a variational Gaussian approximation to the posterior distribution arising from the Poisson model with a Gaussian prior. This is achieved by seeking an optimal Gaussian distribution minimizing the Kullback–Leibler divergence from the posterior distribution to the approximation, or equivalently maximizing the lower bound for the model evidence. We derive an explicit expression for the lower bound, and show the existence and uniqueness of the optimal Gaussian approximation. The lower bound functional can be viewed as a variant of classical Tikhonov regularization that penalizes also the covariance. Then we develop an efficient alternating direction maximization algorithm for solving the optimization problem, and analyze its convergence. We discuss strategies for reducing the computational complexity via low rank structure of the forward operator and the sparsity of the covariance. Further, as an application of the lower bound, we discuss hierarchical Bayesian modeling for selecting the hyperparameter in the prior distribution, and propose a monotonically convergent algorithm for determining the hyperparameter. We present extensive numerical experiments to illustrate the Gaussian approximation and the algorithms.

Keywords: variational Gaussian approximation, Poisson data, hierarchical modeling, Kullback–Leibler divergence, alternating direction maximization

(Some figures may appear in colour only in the online journal)

1. Introduction

This work is concerned with Gaussian approximations to a Poisson noise model for linear inverse problems. The Poisson model is popular for modeling count data, where the response variable follows a Poisson distribution with a parameter that is the exponential of a linear combination of the unknown parameters. The model is especially suitable for low count data, where the standard Gaussian model is inadequate. It has found many successful practical applications, including transmission tomography [12, 41].

One traditional approach to parameter estimation with the Poisson model is the maximum likelihood method or penalized variants with a convex penalty. This leads to a convex optimization problem, whose solution is then taken as an approximation to the true solution. This approach has been extensively studied, and we refer interested readers to the survey [18] for a comprehensive account on important developments along this line. However, this approach gives only a point estimator, and does not allow quantifying the associated uncertainties directly. In this work, we aim at a full Bayesian treatment of the problem, where both the point estimator (mean) and the associated uncertainties (covariance) are of interest [23, 38]. We shall focus on the case of a Gaussian prior, which forms the basis of many other important priors, e.g. sparsity prior via scale mixture representation. Then following the Bayesian procedure, we arrive at a posterior probability distribution, which however is analytically intractable due to the nonstandard form of the likelihood function for the Poisson model. We will explain this more precisely in section 2. To explore the posterior state space, instead of applying popular general-purposed sampling techniques, e.g. Markov chain Monte Carlo (MCMC), we employ a variational Gaussian approximation (VGA). The VGA is one extremely popular approximate inference technique in machine learning [8, 40]. Specifically, we seek an optimal Gaussian approximation to the non-Gaussian posterior distribution with respect to the Kullback–Leibler divergence. The approach leads to a large-scale optimization problem over the mean $\bar{\mathbf{x}}$ and covariance $\mathbf{C}$ (of the Gaussian approximation). In practice, it generally delivers an accurate approximation in an efficient manner, and thus has received immense attention in recent years in many different areas [2, 3, 8, 17]. By its very construction, it also gives a lower bound to the model evidence, which facilitates its use in model selection. However, a systematic theoretical understanding of the approach remains largely missing.

In this work, we shall study the analytical properties and develop an efficient algorithm for the VGA in the context of Poisson data (with the log linear link function). We shall provide a detailed analysis of the resulting optimization problem. The study sheds interesting new insights into the approach from the perspective of regularization. Our main contributions are as follows. First, we derive explicit expressions for the objective functional and its gradient, and establish its strict concavity and the well-posedness of the optimization problem. Second, we develop an efficient numerical algorithm for finding the optimal Gaussian approximation, and discuss its convergence properties. The algorithm is of alternating maximization (coordinate ascent) nature, and it updates the mean $\bar{\mathbf{x}}$ and covariance $\mathbf{C}$ alternatingly by a globally convergent Newton method and a fixed point iteration, respectively. We also discuss strategies for its efficient implementation, by leveraging realistic structure of inverse problems, e.g. low-rank nature of the forward map $\mathbf{A}$ and sparsity of the covariance $\mathbf{C}$, to reduce the computational complexity. Third, we illustrate the use of the evidence lower bound for hyperparameter

selection within a hierarchical Bayesian framework, leading to a purely data-driven approach for determining the regularization parameter, whose proper choice is notoriously challenging. We shall develop a monotonically convergent algorithm for determining the hyperparameter in the Gaussian prior. Last, we illustrate the approach and the algorithms with extensive numerical experiments for one- and two-dimensional examples.

Last, we discuss existing works on Poisson models. The majority of existing works aim at recovering point estimators, either iteratively or by a variational framework [18]. Recently, Bardsley and Luttman [4] described a Metroplis-Hastings algorithm for exploring the posterior distribution (with rectified linear inverse link function), where the proposal samples are drawn from the Laplace approximation (see remark 3.1). The Poisson model (2.2) belongs to generalized linear models (GLMs), to which the VGA has been applied in statistics and machine learning [8, 26, 32, 36]. Ormerod and Wand [32] suggested a variational approximation strategy for fitting GLMs suitable for grouped data. Challis and Barber [8] systematically studied VGA for GLMs and various extensions. The focus of these interesting works [8, 26, 32, 36] is on the development of the general VGA methodology and its applications to concrete problems, and do not study analytical properties and computational techniques for the lower bound functional, which is the main goal of this work.

The rest of the paper is organized as follows. In section 2, we describe the Poisson model, and formulate the posterior probability distribution. Then in section 3, we develop the variational Gaussian approximation, and analyze its basic analytical properties. In section 4, we propose an efficient numerical algorithm for finding the optimal Gaussian approximation, and in section 5, we apply the lower bound to hyperparameter selection within a hierarchical Bayesian framework. In section 6 we present numerical results for several examples. In two appendices, we provide further discussions on the convergence of the fixed point iteration (4.4) and the differentiability of the regularized solution.

2. Notation and problem setting

First we recall some standard notation in linear algebra. Throughout, (real-valued) vectors and matrices are denoted by bold lower- and upper-case letters, respectively, and the vectors are always column vectors. We will use the notation $(\cdot,\cdot)$ to denote the usual Euclidean inner product. We shall slightly abuse the notation $(\cdot,\cdot)$ also for the inner product for matrices. That is, for two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times m}$, we define

$$(\mathbf{X}, \mathbf{Y}) = \mathrm{tr}(\mathbf{X}\mathbf{Y}^t) = \mathrm{tr}(\mathbf{X}^t\mathbf{Y}),$$

where $\mathrm{tr}(\cdot)$ denotes taking the trace of a square matrix, and the superscript t denotes the transpose of a vector or matrix. This inner product induces the usual Frobenius norm for matrices. We shall use extensively the cyclic property of the trace operator $\mathrm{tr}(\cdot)$: for three matrices $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ of appropriate size, there holds

$$\mathrm{tr}(\mathbf{XYZ}) = \mathrm{tr}(\mathbf{YZX}) = \mathrm{tr}(\mathbf{ZXY}).$$

We shall also use the notation $\mathtt{diag}(\cdot)$ for a vector and a square matrix, which gives a diagonal matrix and a column vector from the diagonals of the matrix, respectively, in the same manner as the `diag` function in `MATLAB`. The notation $\mathbb{N} = \{0, 1, \ldots\}$ denotes the set of natural numbers. Further, the notation $\circ$ denotes the Hadamard product of two matrices or vectors. Last, we denote by $\mathcal{S}_m^+ \subset \mathbb{R}^{m\times m}$ the set of symmetric positive definite matrices in $\mathbb{R}^{m\times m}$, $\mathbf{I}_m$ the identity matrix in $\mathbb{R}^{m\times m}$, and by $|\cdot|$ and $\|\cdot\|$ the determinant and the spectral norm, respectively, of a square matrix. Throughout, we view exponential, logarithm and factorial of a vector as componentwise operation.

Next we recall the finite-dimensional Poisson data model. Let $\mathbf{x} \in \mathbb{R}^m$ be the unknown signal, $\mathbf{a}_i \in \mathbb{R}^m$, $i = 1, \ldots, n$, and $\mathbf{y} \in \mathbb{N}^n \subset \mathbb{R}^n$ be the data vector. We stack the column vectors $\mathbf{a}_i$ into a matrix $\mathbf{A}$ by $\mathbf{A} = [\mathbf{a}_i^t] \in \mathbb{R}^{n\times m}$. Given the matrix $\mathbf{A}$ and data $\mathbf{y} \in \mathbb{N}^n$, the Poisson model takes the form:

$$y_i \sim \text{Pois}(\mathrm{e}^{(\mathbf{a}_i,\mathbf{x})}), \quad i = 1, 2, \ldots, n.$$

Thus, the likelihood function $p(y_i|\mathbf{x})$ for the data point y_i is given by

$$p(y_i|\mathbf{x}) = \frac{\lambda_i^{y_i}\mathrm{e}^{-\lambda_i}}{y_i!}, \quad \lambda_i = \mathrm{e}^{(\mathbf{a}_i,\mathbf{x})},\ i = 1, \ldots, n. \tag{2.1}$$

It is worth noting that the exponential function enters into the Poisson parameter λ. This is commonly known as the log link function or log-linear model in the statistical literature [7]. There are several other models for the (inverse) link functions, e.g. rectified-linear and soft-plus [33], each having its own pros and cons for modeling count data. In this work, we shall focus on the log link function. Also this model can be viewed as a simplified statistical model for transmission tomography [12, 41].

The likelihood function $p(y_i|\mathbf{x})$ can be equivalently written as

$$p(y_i|\mathbf{x}) = \mathrm{e}^{y_i(\mathbf{a}_i,\mathbf{x}) - \mathrm{e}^{(\mathbf{a}_i,\mathbf{x})} - \ln(y_i!)}.$$

Under the independent identically distributed (i.i.d.) assumption on the data points y_i, the likelihood function $p(\mathbf{y}|\mathbf{x})$ of the data vector $\mathbf{y}$ is given by

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}) = \mathrm{e}^{(\mathbf{A}\mathbf{x},\mathbf{y}) - (\mathrm{e}^{\mathbf{A}\mathbf{x}},\mathbf{1}_n) - (\ln(\mathbf{y}!),\mathbf{1}_n)}, \tag{2.2}$$

where $\mathbf{1}_n \in \mathbb{R}^n$ is the vector with all entries equal to unity, i.e. $\mathbf{1}_n = [1, \ldots, 1]^t \in \mathbb{R}^n$.

Further, we assume that the unknown $\mathbf{x}$ follows a Gaussian prior $p(\mathbf{x})$, i.e.

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_0, \mathbf{C}_0) := (2\pi)^{-\frac{m}{2}}|\mathbf{C}_0|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^t\mathbf{C}_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)},$$

where $\boldsymbol{\mu}_0 \in \mathbb{R}^m$ and $\mathbf{C}_0 \in \mathcal{S}_m^+$ denote the mean and covariance of the Gaussian prior, respectively, and $\mathcal{N}$ denotes the normal distribution. In the framework of variational regularization, the corresponding penalty $\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_0)^t\mathbf{C}_0^{-1}(\mathbf{x} - \boldsymbol{\mu}_0)$ often imposes certain smoothness constraint. The Gaussian prior $p(\mathbf{x})$ may depend on additional hyperparameters; see section 5 for details. Then by Bayes' formula, the posterior probability distribution $p(\mathbf{x}|\mathbf{y})$ is given by

$$p(\mathbf{x}|\mathbf{y}) = Z^{-1}p(\mathbf{x}, \mathbf{y}), \tag{2.3}$$

where the joint distribution $p(\mathbf{x}, \mathbf{y})$ is defined by

$$p(\mathbf{x}, \mathbf{y}) = (2\pi)^{-\frac{m}{2}}|\mathbf{C}_0|^{-\frac{1}{2}}\mathrm{e}^{(\mathbf{A}\mathbf{x},\mathbf{y}) - (\mathrm{e}^{\mathbf{A}\mathbf{x}},\mathbf{1}_n) - (\ln(\mathbf{y}!),\mathbf{1}_n) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^t\mathbf{C}_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)},$$

and the normalizing constant $Z(\mathbf{y})$, which depends only on the given data $\mathbf{y}$, is given by

$$Z(\mathbf{y}) = p(\mathbf{y}) = \int p(\mathbf{x}, \mathbf{y})\mathrm{d}\mathbf{x}.$$

That is, the normalizing constant Z is an integral living in a very high-dimensional space if the parameter dimension m is large. Thus it is computationally intractable, and so is the posterior distribution $p(\mathbf{x}|\mathbf{y})$, since it also involves the constant Z. The quantity Z is commonly known as

model evidence in the literature, and it underlies many model selection rules, e.g. Bayes factor [24]. Thus the reliable approximation of $Z(\mathbf{y})$ is important in certain tasks.

The posterior distribution $p(\mathbf{x}|\mathbf{y})$ given in (2.3) is the Bayesian solution to the Poisson model (2.1) (under a Gaussian prior), and it contains all the information about the inverse problem. In order to explore the posterior state space, one typically employs Markov chain Monte Carlo methods, which, however, can be prohibitively expensive for high-dimensional problems, apart from the well-known challenge in diagnosing the convergence of the Markov chain. To overcome the challenge, over the last two decades, a large number of approximate inference methods have been developed, including mean-field approximation [40], expectation propagation [30] and variational Gaussian approximation (VGA) [8, 31]. In all these approximations, we aim at finding a best approximate yet tractable distribution $q(\mathbf{x})$ within a family of parametric/nonparametric probability distributions, by minimizing the error in a certain probability metric, prominently the Kullback–Leibler divergence $\mathrm{D}_{\mathrm{KL}}(q\|\,p)$; see section 3.1 below.

In this work, we shall employ the VGA to obtain an optimal Gaussian approximation $q(\mathbf{x})$ to the posterior distribution $p(\mathbf{x}|\mathbf{y})$ in the Kullback–Leibler divergence $\mathrm{D}_{\mathrm{KL}}(q\|\,p)$. Fitting a Gaussian to an intractable distribution is a well-established norm for approximate Bayesian inference, and it has demonstrated success in many practical applications [2, 3, 8, 17]. The popularity can largely be attributed to the fact that the Gaussian approximation is computationally attractive, and yet delivers reasonable accuracy for a wide range of problems, due to the good analytical properties and great flexibility of the Gaussian family. However, analytical properties of approximate inference procedures are rarely studied. In the context of Poisson mixed models, the asymptotic normality of the estimator and its convergence rate was analyzed [16]. In a general setting, some theoretical issues were studied in [29, 34].

3. Gaussian variational approximation

In this section, we recall the Kullback–Leibler divergence, derive explicit expressions for the lower bound functional and its gradient, and discuss basic analytic properties, e.g. concavity and existence.

3.1. Kullback–Leibler divergence

The Kullback–Leibler divergence is one of the most popular metrics for measuring the distance between two probability distributions. The Kullback–Leibler (KL) divergence [28] from one probability distribution p to another distribution q is a functional defined by

$$\mathrm{D}_{\mathrm{KL}}(q\|\,p) = \int q(\mathbf{x})\ln\frac{q(\mathbf{x})}{p(\mathbf{x})}\mathrm{d}\mathbf{x}. \tag{3.1}$$

Clearly, KL divergence is not symmetric and thus not a metric in the mathematical sense. Since the logarithm function $\ln x$ is concave and that q is normalized, i.e. $\int q(\mathbf{x})\mathrm{d}\mathbf{x} = 1$, by Jensen's inequality, we can derive the nonnegativity of the KL divergence:

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(q\|\,p) &= \int q(\mathbf{x})\ln\frac{q(\mathbf{x})}{p(\mathbf{x})}\mathrm{d}\mathbf{x} = -\int q(\mathbf{x})\ln\frac{p(\mathbf{x})}{q(\mathbf{x})}\mathrm{d}\mathbf{x}\\ &\geqslant -\ln\int q(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}\mathrm{d}\mathbf{x} = -\ln\int p(\mathbf{x})\mathrm{d}\mathbf{x} = 0.\end{aligned} \tag{3.2}$$

Further, $\mathrm{D}_{\mathrm{KL}}(q\|\,p) = 0$ if and only if $p = q$ almost everywhere.

Due to unsymmetry of the KL divergence, to find an approximation q to the target distribution p, there are two options, i.e. minimizing either $\mathrm{D}_{\mathrm{KL}}(q\|p)$ or $\mathrm{D}_{\mathrm{KL}}(p\|q)$. These two options lead to different approximations. It was pointed out in [6, section 10.1.2] that minimizing $\mathrm{D}_{\mathrm{KL}}(p\|q)$ tends to find the average of modes of p, while minimizing $\mathrm{D}_{\mathrm{KL}}(q\|p)$ tends to find one exact mode. Traditionally, the former is used in expectation propagation, and the latter in variational Bayes. In this work, we focus on the approach $\min \mathrm{D}_{\mathrm{KL}}(q\|p)$, which leads to the VGA to be described below.

Remark 3.1. In practice, the so-called Laplace approximation is quite popular [39]. Specifically, let $\hat{\mathbf{x}}$ be the maximum *a posteriori* (MAP) estimator $\hat{\mathbf{x}}$, i.e. $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^m} g(\mathbf{x})$, where $g(\mathbf{x}) = -\ln p(\mathbf{x}|\mathbf{y})$ is the negative log posterior distribution. Consider the Taylor expansion of $g(\mathbf{x})$ at the MAP estimator $\hat{\mathbf{x}}$:

$$\begin{aligned} g(\mathbf{x}) &\approx g(\hat{\mathbf{x}}) + (\nabla g(\hat{\mathbf{x}}), \mathbf{x}-\hat{\mathbf{x}}) + \frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^t\mathbf{H}(\mathbf{x}-\hat{\mathbf{x}}) \\ &= g(\hat{\mathbf{x}}) + \frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^t\mathbf{H}(\mathbf{x}-\hat{\mathbf{x}}), \end{aligned}$$

since $\nabla g(\hat{\mathbf{x}})$ vanishes. The Hessian $\mathbf{H}$ of $g(\mathbf{x})$ is given by

$$\mathbf{H} = \mathbf{A}^t\mathrm{diag}(\mathrm{e}^{\mathbf{A}\hat{\mathbf{x}}})\mathbf{A} + \mathbf{C}_0^{-1}.$$

Thus, $\hat{\mathbf{x}}$ might serve as an approximate posterior mean, and the inverse Hessian $\mathbf{H}^{-1}$ as an approximate posterior covariance. However, unlike the VGA discussed below, it lacks the optimality as evidence lower bound (within the Gaussian family), and thus may be suboptimal for model selection etc.

3.2. Variational Gaussian lower bound

Now we derive the variational Gaussian lower bound. By substituting $p(\mathbf{x})$ with the posterior distribution $p(\mathbf{x}|\mathbf{y})$ in (3.1), we obtain

$$\mathrm{D}_{\mathrm{KL}}(q(\mathbf{x})\|\,p(\mathbf{x}|\mathbf{y})) = \int q(\mathbf{x})\ln\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})}\mathrm{d}\mathbf{x}.$$

Since the posterior distribution $p(\mathbf{x}|\mathbf{y})$ depends on the unknown normalizing constant $Z(\mathbf{y})$, the integral on the right hand side is not computable. Nonetheless, given $\mathbf{y}$, $Z(\mathbf{y})$ is fixed. In view of the identity

$$\ln Z = \int q(\mathbf{x})\ln\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x})}\mathrm{d}\mathbf{x} + \int q(\mathbf{x})\ln\frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})}\mathrm{d}\mathbf{x},$$

instead of minimizing $\mathrm{D}_{\mathrm{KL}}(q(\mathbf{x})\|\,p(\mathbf{x}|\mathbf{y}))$, we may equivalently maximize the functional

$$F(q,\mathbf{y}) = \int q(\mathbf{x})\ln\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x})}\mathrm{d}\mathbf{x}. \tag{3.3}$$

By (3.2), we have $\mathrm{D}_{\mathrm{KL}}(q(\mathbf{x})\|\,p(\mathbf{x}|\mathbf{y})) \geqslant 0$, and thus $\ln Z \geqslant F(q,\mathbf{y})$. That is, $F(q,\mathbf{y})$ provides a lower bound on the model evidence Z, for any choice of the distribution q. For any fixed q, $F(q,\mathbf{y})$ may be used as a substitute for the analytically intractable model evidence $Z(\mathbf{y})$, and hence it is called an evidence lower bound (ELBO). Since the data $\mathbf{y}$ is fixed, it will be suppressed from $F(q,\mathbf{y})$ below. In the VGA, we restrict our choice of q to Gaussian distributions.

Meanwhile, a Gaussian distribution $q(\mathbf{x})$ is fully characterized by its mean $\bar{\mathbf{x}} \in \mathbb{R}^m$ and covariance $\mathbf{C} \in \mathcal{S}_m^+ \subset \mathbb{R}^{m\times m}$, i.e.

$$q(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \bar{\mathbf{x}}, \mathbf{C}).$$

Thus, $F(q)$ is actually a function of $\bar{\mathbf{x}} \in \mathbb{R}^m$ and $\mathbf{C} \in \mathcal{S}_m^+$, and will be written as $F(\bar{\mathbf{x}}, \mathbf{C})$ below. Then the approach seeks optimal variational parameters $(\bar{\mathbf{x}}, \mathbf{C})$ to maximize ELBO. This step turns a challenging sampling problem into a computationally more tractable optimization problem.

The next result gives an explicit expression for the lower bound $F(\bar{\mathbf{x}}, \mathbf{C})$.

Proposition 3.1. *For any fixed* $\mathbf{y}, \boldsymbol{\mu}_0$ *and* $\mathbf{C}_0$*, the lower bound* $F(\bar{\mathbf{x}}, \mathbf{C})$ *is given by*

$$\begin{aligned} F(\bar{\mathbf{x}}, \mathbf{C}) = (\mathbf{y}, \mathbf{A}\bar{\mathbf{x}}) - (\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)}) - \frac{1}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^t \mathbf{C}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \\ - \frac{1}{2}\mathrm{tr}(\mathbf{C}_0^{-1}\mathbf{C}) + \frac{1}{2}\ln|\mathbf{C}| - \frac{1}{2}\ln|\mathbf{C}_\mathbf{0}| + \frac{m}{2} - (\mathbf{1}_n, \ln(\mathbf{y}!)). \end{aligned} \tag{3.4}$$

Proof. By the definition of the functional $F(\bar{\mathbf{x}}, \mathbf{C})$ and the joint distribution $p(\mathbf{x}, \mathbf{y})$, we have

$$\begin{aligned} F(\bar{\mathbf{x}}, \mathbf{C}) = \int \mathcal{N}(\mathbf{x}; \bar{\mathbf{x}}, \mathbf{C}) \Big[\ln|\mathbf{C}_0|^{-\frac{1}{2}} - \ln|\mathbf{C}|^{-\frac{1}{2}} + (\mathbf{A}\mathbf{x}, \mathbf{y}) - (\mathrm{e}^{\mathbf{A}\mathbf{x}}, \mathbf{1}_n) - (\ln(\mathbf{y}!), \mathbf{1}_n) \\ - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_0)^t \mathbf{C}_0^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^t \mathbf{C}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \Big] \mathrm{d}\mathbf{x}. \end{aligned}$$

It suffices to evaluate the integrals termwise. Clearly, we have $\int \mathcal{N}(\mathbf{x}; \bar{\mathbf{x}}, \mathbf{C})(\mathbf{A}\mathbf{x}, \mathbf{y})\mathrm{d}\mathbf{x} = (\mathbf{A}\bar{\mathbf{x}}, \mathbf{y})$. Next, using moment generating function, we have

$$\begin{aligned} \int \mathcal{N}(\mathbf{x}; \bar{\mathbf{x}}, \mathbf{C})(\mathrm{e}^{\mathbf{A}\mathbf{x}}, \mathbf{1}_n)\mathrm{d}\mathbf{x} &= \sum_i \int \mathcal{N}(\mathbf{x}; \bar{\mathbf{x}}, \mathbf{C}) \mathrm{e}^{(\mathbf{a}_i, \mathbf{x})} \mathrm{d}\mathbf{x} \\ &= \sum_i \mathrm{e}^{(\mathbf{a}_i, \bar{\mathbf{x}}) + \frac{1}{2}\mathbf{a}_i^t \mathbf{C}\mathbf{a}_i} = (\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)}). \end{aligned}$$

With the Cholesky decomposition $\mathbf{C} = \mathbf{L}\mathbf{L}^t$, for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, $\mathbf{x} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z} \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{C})$. This and the bias-variance decomposition yield ($\mathbb{E}_{q(\mathbf{x})}[\cdot]$ takes expectation with respect to the density $q(\mathbf{x})$): for any symmetric $\mathrm{X} \in \mathbb{R}^{mxm}$

$$\mathbb{E}_{q(\mathbf{x})}[\mathbf{x}^t \mathrm{X} \mathbf{x}] = \mathbb{E}_{\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}_m)}[(\boldsymbol{\mu} + \mathbf{L}\mathbf{z})^t \mathrm{X}(\boldsymbol{\mu} + \mathbf{L}\mathbf{z})] = \boldsymbol{\mu}^t \mathrm{X}\boldsymbol{\mu} + \mathbb{E}_{\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}_m)}[\mathbf{z}^t \mathbf{L}^t \mathrm{X}\mathbf{L}\mathbf{z}].$$

By the cyclic property of trace, we have $\mathbb{E}_{\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}_m)}[\mathbf{z}^t \mathbf{L}^t \mathrm{X}\mathbf{L}\mathbf{z}] = \mathrm{tr}(\mathbf{L}^t \mathrm{X}\mathbf{L}) = \mathrm{tr}(\mathrm{X}\mathbf{L}\mathbf{L}^t) = \mathrm{tr}(\mathrm{X}\mathbf{C})$. In particular, this gives

$$\mathbb{E}_{q(\mathbf{x})}[(\mathbf{x} - \boldsymbol{\mu}_0)^t \mathbf{C}_0^{-1}(\mathbf{x} - \boldsymbol{\mu}_0)] = (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^t \mathbf{C}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) + \mathrm{tr}(\mathbf{C}_0^{-1}\mathbf{C}),$$

and

$$\mathbb{E}_{q(\mathbf{x})}[(\mathbf{x} - \bar{\mathbf{x}})^t \mathbf{C}^{-1}(\mathbf{x} - \bar{\mathbf{x}})] = m.$$

Collecting preceding identities completes the proof of the proposition. □

Remark 3.2. The terms in the functional $F(\bar{\mathbf{x}}, \mathbf{C})$ in (3.4) admit interesting interpretation in the lens of classical Tikhonov regularization (see, e.g. [11, 19, 37]). To this end, it is instructive to rewrite it as

$$\begin{aligned} F(\bar{\mathbf{x}}, \mathbf{C}) =& (\mathbf{y}, \mathbf{A}\bar{\mathbf{x}}) - (\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)}) - (\mathbf{1}_n, \ln(\mathbf{y}!)) \\ & - \frac{1}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^t \mathbf{C}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \\ & - \frac{1}{2}\mathrm{tr}(\mathbf{C}_0^{-1}\mathbf{C}) + \frac{1}{2}\ln|\mathbf{C}| - \frac{1}{2}\ln|\mathbf{C}_0| + \frac{m}{2}. \end{aligned}$$

The first line represents the fidelity or 'pseudo-likelihood' function. It is worth noting that it actually involves the covariance $\mathbf{C}$. In the absence of the covariance $\mathbf{C}$, it recovers the familiar log likelihood for Poisson data, see remark 3.1. The second line imposes a quadratic penalty on the mean $\bar{\mathbf{x}}$. This term recovers the familiar penalty in Tikhonov regularization (except that it is imposed on $\bar{\mathbf{x}}$). Recall that the function $-\ln|\mathbf{C}|$ is strictly convex in $\mathbf{C} \in \mathcal{S}_m^+$ [13, lemma 6.2.2]. Thus, one may define the corresponding Bregman divergence $d(\mathbf{C}, \mathbf{C}_0)$. In view of the identities [9]

$$\frac{\partial}{\partial \mathbf{C}}\mathrm{tr}(\mathbf{C}\mathbf{C}_0^{-1}) = \mathbf{C}_0^{-1} \quad \text{and} \quad \frac{\partial}{\partial \mathbf{C}}\ln|\mathbf{C}| = \mathbf{C}^{-1} \tag{3.5}$$

simple computation gives the following expression for the divergence:

$$d(\mathbf{C}, \mathbf{C}_0) = \mathrm{tr}(\mathbf{C}_0^{-1}\mathbf{C}) - \ln|\mathbf{C}_0^{-1}\mathbf{C}| - m \geqslant 0.$$

Statistically, it is the Kullback–Leibler divergence between two Gaussians of identical mean. The divergence $d(\mathbf{C}, \mathbf{C}_0)$ provides a distance measure between the prior covariance $\mathbf{C}_0$ and the posterior one $\mathbf{C}$. Let $\{(\lambda_i, \mathbf{v}_i)\}_{i=1}^m$ be the pairs of generalized eigenvalues and eigenfunctions of the pencil $(\mathbf{C}, \mathbf{C}_0)$, i.e. $\mathbf{C}\mathbf{v}_i = \lambda_i \mathbf{C}_0 \mathbf{v}_i$. Then it can be expressed as

$$d(\mathbf{C}, \mathbf{C}_0) = \sum_{i=1}^{m} (\lambda_i - \ln\lambda_i - 1).$$

This identity directly indicates that $d(\mathbf{C}, \mathbf{C}_0) \leqslant c$ implies $\|\mathbf{C}\| \leqslant c$ and $\|\mathbf{C}^{-1}\| \leqslant c$, where here and below c denotes a generic constant which may change at each occurrence.

Thus, the third line regularizes the posterior covariance $\mathbf{C}$ by requesting nearness to the prior one $\mathbf{C}_0$ in Bregman divergence. It is interesting to observe that the Gaussian prior implicitly induces a penalty on $\mathbf{C}$, although it is not directly enforced. In statistics, the Bregman divergence $d(\mathbf{C}, \mathbf{C}_0)$ is also known as Stein's loss [21]. In recent years, the Bregman divergence $d(\mathbf{C}, \mathbf{C}_0)$ has been employed in clustering, graphical models, sparse covariance estimate and low-rank matrix recovery etc [27, 35].

3.3. Theoretical properties of the lower bound

Now we study basic analytical properties, i.e. concavity, existence and uniqueness of maximizer, and gradient of the functional $F(\bar{\mathbf{x}}, \mathbf{C})$ defined in (3.4), from the perspective of optimization.

A first result shows the concavity of $F(\bar{\mathbf{x}}, \mathbf{C})$. Let X and Y be two convex sets. Recall that a functional $f : X \times Y \to \mathbb{R}$ is said to be jointly concave, if and only if

$$f(\lambda x_1 + (1-\lambda)x_2, \lambda y_1 + (1-\lambda)y_2) \geqslant \lambda f(x_1, y_1) + (1-\lambda) f(x_2, y_2)$$

for all $x_1, x_2 \in X,\ y_1, y_2 \in Y$ and $\lambda \in [0,1]$. Further, f is called strictly jointly concave if the inequality is strict for any $(x_1, y_1) \neq (x_1, y_1)$ and $\lambda \in (0,1)$. Clearly, $\mathcal{S}_m^+$ is a convex set.

Theorem 3.1. *For any $\mathbf{C}_0 \in \mathcal{S}_m^+$, the functional $F(\bar{\mathbf{x}}, \mathbf{C})$ is strictly jointly concave with respect to $\bar{\mathbf{x}} \in \mathbb{R}^m$ and $\mathbf{C} \in \mathcal{S}_m^+$.*

Proof. It suffices to consider the terms apart from the linear terms $(\mathbf{y}, \mathbf{A}\bar{\mathbf{x}})$ and $-\frac{1}{2}\mathrm{tr}(\mathbf{C}_0^{-1}\mathbf{C})$ and the constant term $-\frac{1}{2}\ln|\mathbf{C_0}| + \frac{m}{2} - (\mathbf{1}_n, \ln(\mathbf{y}!))$. Since $\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)$ is linear in $\bar{\mathbf{x}}$ and $\mathbf{C}$, and exponentiation preserves convexity, the term $-(\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})$ is also jointly concave. Next, the term $-\frac{1}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^t\mathbf{C}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$ is strictly concave for any $\mathbf{C}_0 \in \mathcal{S}_m^+$. Last, the log-determinant $\ln|\mathbf{C}|$ is strictly concave over $\mathcal{S}_m^+$ [13, lemma 6.2.2]. The assertion follows since strict concavity is preserved under summation. □

Next, we show the well-posedness of the optimization problem in VGA.

Theorem 3.2. *There exists a unique pair of $(\bar{\mathbf{x}}, \mathbf{C})$ solving the optimization problem*

$$\max_{\bar{\mathbf{x}} \in \mathbb{R}^m, \mathbf{C} \in \mathcal{S}_m^+} F(\bar{\mathbf{x}}, \mathbf{C}) \tag{3.6}$$

Proof. The proof follows by direct methods in calculus of variation, and we only briefly sketch it. Clearly, there exists a maximizing sequence, denoted by $\{(\bar{\mathbf{x}}^k, \mathbf{C}^k)\} \subset \mathbb{R}^m \times \mathcal{S}_m^+$, and we may assume $F(\bar{\mathbf{x}}^k, \mathbf{C}^k) \geqslant c =: F(\boldsymbol{\mu}_0, \mathbf{C}_0)$. Thus, by (3.4) in proposition 3.1 and the divergence $d(\mathbf{C}, \mathbf{C}_0)$, we have

$$(\mathbf{A}\bar{\mathbf{x}}^k, \mathbf{y}) - (\bar{\mathbf{x}}^k - \boldsymbol{\mu}_0)^t\mathbf{C}_0^{-1}(\bar{\mathbf{x}}^k - \boldsymbol{\mu}_0) - d(\mathbf{C}^k, \mathbf{C}_0) \geqslant c + (\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}^k + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}^k\mathbf{A}^t)}, \mathbf{1}_n) \geqslant c.$$

By the Cauchy–Schwarz inequality, we have $(\bar{\mathbf{x}}^k - \boldsymbol{\mu}_0)^t\mathbf{C}_0^{-1}(\bar{\mathbf{x}}^k - \boldsymbol{\mu}_0) + d(\mathbf{C}^k, \mathbf{C}_0) \leqslant c$. This immediately implies a uniform bound on $\{(\bar{\mathbf{x}}^k, \mathbf{C}^k)\}$ and $\{(\mathbf{C}^k)^{-1}\}$. Thus, there exists a convergent subsequence, relabeled as $\{(\bar{\mathbf{x}}^k, \mathbf{C}^k)\}$, with a limit $(\bar{\mathbf{x}}^*, \mathbf{C}^*) \in \mathbb{R}^m \times \mathcal{S}_m^+$. Then by the continuity of the functional F in $(\bar{\mathbf{x}}, \mathbf{C})$, we deduce that $(\bar{\mathbf{x}}^*, \mathbf{C}^*)$ is a maximizer to $F(\bar{\mathbf{x}}, \mathbf{C})$, i.e. the existence of a maximizer. The uniqueness follows from the strict joint-concavity of $F(\bar{\mathbf{x}}, \mathbf{C})$, see theorem 3.1. □

Since F is composed of smooth functions, clearly it is smooth. Next we give the gradient formulae, which are useful for developing numerical algorithms below.

Theorem 3.3. *The gradients of the functional $F(\bar{\mathbf{x}}, \mathbf{C})$ with respect to $\bar{\mathbf{x}}$ and $\mathbf{C}$ are respectively given by*

$$\frac{\partial F}{\partial \bar{\mathbf{x}}} = \mathbf{A}^t\mathbf{y} - \mathbf{A}^t\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)} - \mathbf{C}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0),$$
$$\frac{\partial F}{\partial \mathbf{C}} = \frac{1}{2}[-\mathbf{A}^t\mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})\mathbf{A} - \mathbf{C}_0^{-1} + \mathbf{C}^{-1}].$$

Proof. Let $\mathbf{d} = \mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)$. Then by the chain rule

$$\frac{\partial}{\partial \bar{x}_i}(\mathbf{1}_n, \mathrm{e}^{\mathbf{d}}) = \frac{\partial}{\partial \bar{x}_i}\sum_{j=1}^{n}\mathrm{e}^{d_j} = \sum_{j=1}^{n}\frac{\partial \mathrm{e}^{d_j}}{\partial d_j}\frac{\partial d_j}{\partial \bar{x}_i} = \sum_{j=1}^{n}\mathrm{e}^{d_j}(A)_{ji}.$$

That is, we have $\frac{\partial}{\partial \bar{\mathbf{x}}}(\mathbf{1}_n, \mathrm{e}^{\mathbf{d}}) = \mathbf{A}^t\mathrm{e}^{\mathbf{d}}$, showing the first formula. Next we derive the gradient with respect to the covariance $\mathbf{C}$. In view of (3.5), it remains to differentiate the term $(\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})$ with respect to $\mathbf{C}$. To this end, let $\mathbf{H}$ be a small perturbation to $\mathbf{C}$. By Taylor expansion, and with the diagonal matrix $\mathbf{D} = \mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})$, we deduce

$$(\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{A}(\mathbf{C}+\mathbf{H})\mathbf{A}^t)}) - (\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)}) = (\mathbf{D}, \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{H}\mathbf{A}^t)) + \mathcal{O}(\|\mathbf{H}\|^2).$$

Since the matrix $\mathbf{D}$ is diagonal, by the cyclic property of trace, we have

$$(\mathbf{D}, \mathrm{diag}(\mathbf{A}\mathbf{H}\mathbf{A}^t)) = (\mathbf{D}, \mathbf{A}\mathbf{H}\mathbf{A}^t) = \mathrm{tr}(\mathbf{D}\mathbf{A}\mathbf{H}^t\mathbf{A}^t) = \mathrm{tr}(\mathbf{A}^t\mathbf{D}\mathbf{A}\mathbf{H}^t) = (\mathbf{A}^t\mathbf{D}\mathbf{A}, \mathbf{H}).$$

Now the definition of the gradient completes the proof. □

An immediate corollary is the following optimality system.

Corollary 3.1. *The necessary and sufficient optimality system of problem (3.6) is given by*

$$\mathbf{A}^t\mathbf{y} - \mathbf{A}^t\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)} - \mathbf{C}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) = 0,$$
$$\mathbf{C}^{-1} - \mathbf{A}^t\mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})\mathbf{A} - \mathbf{C}_0^{-1} = 0.$$

Remark 3.3. Challis and Barber [8] showed that for log-concave site posterior potentials, the variational lower bound is jointly concave in $\bar{\mathbf{x}}$ and the Cholesky factor $\mathbf{L}$ of the covariance $\mathbf{C}$. This assertion holds also for the lower bound $F(\bar{\mathbf{x}}, \mathbf{C})$ in (3.4), i.e. joint concavity with respect to $(\bar{\mathbf{x}}, \mathbf{L})$.

Remark 3.4. Corollary 3.1 indicates that the covariance $\mathbf{C}^*$ of the optimal Gaussian approximation $q^*(\mathbf{x})$ is of the following form:

$$(\mathbf{C}^*)^{-1} = \mathbf{C}_0^{-1} + \mathbf{A}^t\mathbf{D}\mathbf{A},$$

for some diagonal matrix $\mathbf{D}$. Thus it is tempting that one may minimize with respect to $\mathbf{D}$ instead of $\mathbf{C}$ in order to reduce the complexity of the algorithm, by reducing the number of unknowns from m^2 to m. However, F is generally not concave with respect to $\mathbf{D}$; see [26] for a one-dimensional counterexample. The loss of concavity might complicate the analysis and computation.

Remark 3.5. In practice, the parameter $\mathbf{x}$ in the model (2.2) often admits physical constraint. Thus it is natural to impose a box constraint on the mean $\bar{\mathbf{x}}$ in problem (3.6), e.g. $c_l \leqslant \bar{x}_i \leqslant c_u$, $i = 1, \ldots, m$, for some $c_l < c_u$. This can be easily incorporated into the optimality system in corollary 3.1, and the algorithms below remain valid upon minor changes, e.g. including a pointwise projection operator in the update of $\bar{\mathbf{x}}$.

4. Numerical algorithm and its complexity analysis

Now we develop an efficient numerical algorithm, which is of alternating direction maximization type, provide an analysis of its complexity, and discuss strategies for complexity reduction.

4.1. Numerical algorithm

In view of the strict concavity of $F(\bar{\mathbf{x}}, \mathbf{C})$, it suffices to solve the optimality system (see corollary 3.1):

$$\mathbf{A}^t\mathbf{y} - \mathbf{A}^t e^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)} - \mathbf{C}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) = 0, \tag{4.1}$$

$$\mathbf{C}^{-1} - \mathbf{A}^t \mathrm{diag}(e^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})\mathbf{A} - \mathbf{C}_0^{-1} = 0. \tag{4.2}$$

This consists of a coupled nonlinear system for $(\bar{\mathbf{x}}, \mathbf{C})$. We shall solve the system by alternatingly maximizing $F(\bar{\mathbf{x}}, \mathbf{C})$ with respect to $\bar{\mathbf{x}}$ and $\mathbf{C}$, i.e. coordinate ascent. From the strict concavity in theorem 3.1, we deduce that for a fixed $\mathbf{C}$, (4.1) has a unique solution $\bar{\mathbf{x}}$, and similarly, for a fixed $\bar{\mathbf{x}}$, (4.2) has a unique solution $\mathbf{C}$. Below, we discuss the efficient numerical solution of (4.1) and (4.2).

4.1.1. Newton method for updating $\bar{\mathbf{x}}$. To solve $\bar{\mathbf{x}}$ from (4.1), for a fixed $\mathbf{C}$, we employ a Newton method. Let the nonlinear map $\mathbf{G} : \mathbb{R}^m \to \mathbb{R}^m$ be defined by

$$\mathbf{G}(\bar{\mathbf{x}}) = \mathbf{A}^t e^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)} + \mathbf{C}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) - \mathbf{A}^t\mathbf{y}.$$

The Jacobian $\partial\mathbf{G}$ of the map $\mathbf{G}$ is given by

$$\partial\mathbf{G}(\bar{\mathbf{x}}) = \mathbf{A}^t \mathrm{diag}(e^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})\mathbf{A} + \mathbf{C}_0^{-1} \geqslant \mathbf{C}_0^{-1},$$

where the partial ordering $\geqslant$ is in the sense of symmetric positive definite matrix, i.e. $\mathbf{X} \geqslant \mathbf{Y}$ if and only if $\mathbf{X} - \mathbf{Y}$ is positive semidefinite. That is, the Jacobian $\partial\mathbf{G}(\bar{\mathbf{x}})$ is uniformly invertible (since the prior covariance $\mathbf{C}_0^{-1}$ is invertible). This concurs with the strict concavity of the functional $F(\bar{\mathbf{x}}, \mathbf{C})$ in $\bar{\mathbf{x}}$.

This motivates the use of the Newton method or its variants: for a nonlinear system with uniformly invertible Jacobians, the Newton method converges globally [25]. Specifically, given $\bar{\mathbf{x}}^0$, we iterate

$$\partial\mathbf{G}(\bar{\mathbf{x}}^k)\delta\bar{\mathbf{x}} = -\mathbf{G}(\bar{\mathbf{x}}^k), \qquad \bar{\mathbf{x}}^{k+1} = \bar{\mathbf{x}}^k + \delta\bar{\mathbf{x}}. \tag{4.3}$$

The main cost of the Newton update (4.3) lies in solving the linear system involving $\partial\mathbf{G}(\bar{\mathbf{x}}^k)$. Clearly, the Jacobian $\partial\mathbf{G}(\bar{\mathbf{x}}^k)$ is symmetric and positive definite, and thus the (preconditioned) conjugate gradient method is a natural choice for solving the linear system. One may use $\mathbf{C}_0^{-1}$ (or the diagonal part of the Jacobian $\partial\mathbf{G}(\bar{\mathbf{x}})$) as a preconditioner. It is worth noting that inverting the Jacobian $\partial\mathbf{G}(\bar{\mathbf{x}})$ is identical with one fixed point update of the covariance $\mathbf{C}$ below. In the presence of *a priori* structural information, this can be carried out efficiently even for very large-scale problems; see section 4.2 below for further details. Due to the fast local convergence of the Newton method, a few iterations suffice the desired accuracy, which is fully confirmed by our numerical experiments.

4.1.2. Fixed-point method for updating $\mathbf{C}$. Next we turn to the solution of (4.2) for updating $\mathbf{C}$, with $\bar{\mathbf{x}}$ fixed. There are several different strategies, and we shall describe two of them below. The first option is to employ a Newton method. Let the nonlinear map $\mathbf{S} : \mathbb{R}^{m\times m} \to \mathbb{R}^{m\times m}$ be defined by

$$\mathbf{S}(\mathbf{C}) = \mathbf{C}^{-1} - \mathbf{C}_0^{-1} - \mathbf{A}^t \mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{x}+\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})\mathbf{A}.$$

The Jacobian $\partial\mathbf{S}$ of the map $\mathbf{S}$ is given by

$$\partial\mathbf{S}(\mathbf{C})[\mathbf{H}] = -\mathbf{C}^{-1}\mathbf{H}\mathbf{C}^{-1} - \mathbf{A}^t \mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})\mathrm{diag}(\mathbf{A}\mathbf{H}\mathbf{A}^t)\mathbf{A}.$$

It can be verified that the map $\partial\mathbf{S}(\mathbf{C})$ is symmetric with a uniformly bounded inverse (see the proof of theorem B.1 in the appendix for details). However, its explicit form seems not available due to the presence of the operator diag. Nonetheless, one can apply a (preconditioned) conjugate gradient method for updating $\mathbf{C}$. The Newton iteration is guaranteed to converge globally.

The second option is to use a fixed-point iteration. This choice is very attractive since it avoids solving huge linear systems. Specifically, given an initial guess $\mathbf{C}^0$, we iterate by

$$\mathbf{D}^k = \mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}^k\mathbf{A}^t)}), \qquad \mathbf{C}^{k+1} = (\mathbf{C}_0^{-1} + \mathbf{A}^t\mathbf{D}^k\mathbf{A})^{-1}. \tag{4.4}$$

Conceptually, it has the flavor of a classical fixed point scheme for solving algebraic Riccati equations in Kalman filtering [1], and it has also been used in a slightly different context of variational inference with Gaussian processes [26]. Numerically, each inner iteration of (4.4) involves computing the vector $\mathrm{diag}(\mathbf{A}\mathbf{C}^k\mathbf{A}^t)$ (which should be regarded as computing $\mathbf{a}_i\mathbf{C}^k\mathbf{a}_i^t$, $i = 1, \ldots, m$, instead of forming the full matrix $\mathbf{A}\mathbf{C}^k\mathbf{A}^t$) and a matrix inversion.

Next we briefly discuss the convergence of (4.4). Clearly, for all iterates $\mathbf{C}^k$, we have $\mathbf{C}^k \leqslant \mathbf{C}_0$. We claim $\lambda_{\max}(\mathbf{C}^k) \leqslant \lambda_{\max}(\mathbf{C}_0)$. To see this, let $\mathbf{v} \in \mathbb{R}^m$ be a unit eigenvector corresponding to the largest eigenvalue $\lambda_{\max}(\mathbf{C}^k)$, i.e. $\mathbf{v}^t\mathbf{C}^k\mathbf{v} = \lambda_{\max}(\mathbf{C}^k)$. Then by the min-max principle

$$\lambda_{\max}(\mathbf{C}^k) = \mathbf{v}^t\mathbf{C}^k\mathbf{v} \leqslant \mathbf{v}^t\mathbf{C}_0\mathbf{v} \leqslant \sup_{\mathbf{v}\in\mathbb{S}^m} \mathbf{v}^t\mathbf{C}_0\mathbf{v} = \lambda_{\max}(\mathbf{C}_0).$$

Thus, the sequence $\{\mathbf{C}^k\}_{k=1}^{\infty}$ generated by the iteration (4.4) is uniformly bounded in the spectral norm (and thus any norm due to the norm equivalence in a finite-dimensional space). Hence, there exists a convergent subsequence, also relabeled as $\{\mathbf{C}^k\}$, such that $\mathbf{C}^k \to \mathbf{C}^*$, for some $\mathbf{C}^*$. In practice, the iterates converge fairly steadily to the unique solution to (4.2), which however remains to be established. In appendix A, we show a certain 'monotone' type convergence of (4.4) for the initial guess $\mathbf{C}^0 = \mathbf{C}_0$.

4.1.3. Variational Gaussian approximation algorithm. With the preceding two inner solvers, we are ready to state the complete procedure in algorithm 1. One natural stopping criterion at Step 7 is to monitor ELBO. However, computing ELBO can be expensive and cheap alternatives, e.g. relative change of the mean $\bar{\mathbf{x}}$, might be considered. Note that Step 3 of algorithm 1, i.e. randomized singular value decomposition (rSVD), has to be carried out only once, and it constitutes a preprocessing step. Its crucial role will be discussed in section 4.2 below.

With exact inner updates $(\bar{\mathbf{x}}^k, \mathbf{C}^k)$, by the alternating maximizing property, the sequence $\{F(\bar{\mathbf{x}}^k, \mathbf{C}^k)\}$ is guaranteed to be monotonically increasing, i.e.

$$F(\bar{\mathbf{x}}^0, \mathbf{C}^0) \leqslant F(\bar{\mathbf{x}}^1, \mathbf{C}^0) \leqslant F(\bar{\mathbf{x}}^1, \mathbf{C}^1) \leqslant \ldots \leqslant F(\bar{\mathbf{x}}^k, \mathbf{C}^k) \leqslant \ldots,$$

with the inequality being strict until convergence is reached. Further, $F(\bar{\mathbf{x}}^k, \mathbf{C}^k) \leqslant \ln Z(\mathbf{y})$. Thus, $\{F(\bar{\mathbf{x}}^k, \mathbf{C}^k)\}$ converges. Further, by [5, proposition 2.7.1], the coordinate ascent method converges if the maximization with respect to each coordinate is uniquely attained. Clearly, algorithm 1 is a coordinate ascent method for $F(\bar{\mathbf{x}}, \mathbf{C})$, and $F(\bar{\mathbf{x}}, \mathbf{C})$ satisfies the unique solvability condition. Thus the sequence $\{(\bar{\mathbf{x}}^k, \mathbf{C}^k)\}$ generated by algorithm 1 converges to the unique maximizer of $F(\bar{\mathbf{x}}, \mathbf{C})$.

Algorithm 1. Variational Gaussian approximation algorithm.

1: Input: $(\mathbf{A}, \mathbf{y})$, specify the prior $(\boldsymbol{\mu}_0, \mathbf{C}_0)$, and the maximum number K of iterations
2: Initialize $\bar{\mathbf{x}} = \bar{\mathbf{x}}^1$ and $\mathbf{C} = \mathbf{C}^1$;
3: SVD: $(\mathbf{U}, \Sigma, \mathbf{V}) = \mathrm{rSVD}(\mathbf{A})$;
4: **for** $k = 1, 2, \ldots, K$ **do**
5: Update the mean $\bar{\mathbf{x}}^{k+1}$ by (4.3);
6: Update the covariance $\mathbf{C}^{k+1}$ by (4.4);
7: Check the stopping criterion.
8: **end for**
9: Output: $(\bar{\mathbf{x}}, \mathbf{C})$

4.2. Complexity analysis and reduction

Now we analyze the computational complexity of algorithm 1, and describe strategies for complexity reduction, in order to arrive at a scalable implementation. When evaluating the functional $F(\bar{\mathbf{x}}, \mathbf{C})$ and its gradient, the constant terms can be precomputed. Thus, it suffices to analyze the terms that will be updated. Standard linear algebra [14] gives the following operational complexity.

- The complexity of evaluating the objective functional $F(\bar{\mathbf{x}}, \mathbf{C})$ is $\mathcal{O}(m^2 n + m^3)$:
  - the inner product $-(\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)}) \sim \mathcal{O}(m^2 n)$
  - the matrix determinant $\ln|\mathbf{C}| \sim \mathcal{O}(m^3)$
- The complexity of evaluating the gradient $\frac{\partial F}{\partial \bar{\mathbf{x}}}$ is $\mathcal{O}(m^2 n)$:
  - the matrix-vector product $\mathbf{A}^t \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)} \sim \mathcal{O}(m^2 n)$
- The complexity of evaluating the gradient $\frac{\partial F}{\partial \mathbf{C}}$ is $\mathcal{O}(m^2 n + m^3)$:
  - the matrix product $\mathbf{A}^t \mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})\mathbf{A} \sim \mathcal{O}(m^2 n)$
  - the matrix inversion $\mathbf{C}^{-1} \sim \mathcal{O}(m^3)$.

In summary, evaluating ELBO $F(\bar{\mathbf{x}}, \mathbf{C})$ and its gradients each involves $\mathcal{O}(nm^2 + m^3)$ complexity, which is infeasible for large-scale problems. The most expensive piece lies in matrix products/inversion, e.g. $(\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})$, $\mathbf{A}^t \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)}$ and $\mathbf{A}^t \mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})\mathbf{A}$. The log-determinant $\ln|\mathbf{C}|$ can be approximated accurately with $\mathcal{O}(m^2)$ operations by a stochastic algorithm [42]. In many practical inverse problems, there do exist structures: (i) $\mathbf{A}$ is low rank, and (ii) $\mathbf{C}$ is sparse, which can be leveraged to reduce the per-iteration cost.

First, for many inverse problems, the matrix $\mathbf{A}$ is ill-conditioned, and the singular values decay to zero. Thus, $\mathbf{A}$ naturally has a low-rank structure. The effective rank r is determined by the decay rate of the singular values. In this work, we assume a known rank r. The rSVD is a powerful technique for obtaining low-rank approximations [15]. For a rank r matrix, the rSVD can yield an accurate approximation with $\mathcal{O}(mn \ln r + (m+n)r^2)$ operations [15, p 225]. We

denote the rSVD approximation by $\mathbf{A} \approx \mathbf{U}\Sigma\mathbf{V}^t$, where the matrices $\mathbf{U} \in \mathbb{R}^{n\times r}$ and $\mathbf{V} \in \mathbb{R}^{m\times r}$ are column orthonormal, and $\Sigma \in \mathbb{R}^{r\times r}$ is diagonal with its entries ordered nonincreasingly.

Second, the covariance $\mathbf{C}$ is approximately sparse, and each row/column has at most s nonzero entries. This reflects the fact that only (physically) neighboring elements are highly correlated, and there is no long range correlation. This choice will be implemented in the numerical experiments for 2D image deblurring. Naturally, one can also consider a more flexible option by adaptively selecting the sparsity pattern. This can be achieved by penalizing of the off-diagonal entries of $\mathbf{C}$ by the ℓ^1-norm, which allows automatically detecting significant correlation [35]. Other structures, e.g. low-rank plus sparsity, offer potential alternatives. We leave these advanced options to a future study.

Under these structural assumptions, the complexity of computing the terms $(\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{ACA}^t)})$, $\mathbf{A}^t\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{ACA}^t)}$ and $\mathbf{A}^t\mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{ACA}^t)})\mathbf{A}$ can be reduced to $\mathcal{O}(smn)$. Thus, the complexity of calculating F and $\frac{\partial F}{\partial \bar{\mathbf{x}}}$ is reduced to $\mathcal{O}(smn + m^2)$. For the matrix inversion in (4.4), we exploit the low-rank structure of $\mathbf{A}$. Upon recalling the low-rank approximation of $\mathbf{A}$ and the Sherman-Morrison-Woodbury formula [14, p 65], i.e.

$$(\tilde{\mathbf{A}} + \tilde{\mathbf{U}}\tilde{\mathbf{V}})^{-1} = \tilde{\mathbf{A}}^{-1} - \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{U}}(\mathbf{I} + \tilde{\mathbf{V}}\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{U}})^{-1}\tilde{\mathbf{V}}\tilde{\mathbf{A}}^{-1},$$

we deduce (with $\mathbf{D} = \mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}+\frac{1}{2}\mathrm{diag}(\mathbf{ACA}^t)})$)

$$\mathbf{C} = \mathbf{C}_0 - \mathbf{C}_0\mathbf{V}\Sigma\mathbf{U}^t\mathbf{D}\mathbf{U}\Sigma(\mathbf{I} + \mathbf{V}^t\mathbf{C}_0\mathbf{V}\Sigma\mathbf{U}^t\mathbf{D}\mathbf{U}\Sigma)^{-1}\mathbf{V}^t\mathbf{C}_0. \tag{4.5}$$

Note that the inversion step only involves a matrix in $\mathbb{R}^{r\times r}$, and can be carried out efficiently. The sparsity structure on $\mathbf{C}$ can be enforced by computing only the respective entries. Then the update formula (4.5) can be achieved in $\mathcal{O}(smn + r^2n + r^2m)$ operations. In comparison with the $\mathcal{O}(m^3 + nm^2)$ complexity of the direct implementation, this represents a substantial complexity reduction.

5. Hyperparameter choice with hierarchical model

When encoding prior knowledge about the unknown $\mathbf{x}$ into the prior $p(\mathbf{x})$, it is often necessary to tune its strength, a scalar parameter commonly known as hyperparameter. It plays the role of the regularization parameter in variational regularization [19, chapter 7], where its proper choice is notoriously challenging. In the Gaussian prior $p(\mathbf{x})$, $\mathbf{C}_0 = \alpha^{-1}\bar{\mathbf{C}}_0$, where $\bar{\mathbf{C}}_0$ describes the interaction structure and the scalar α determines the strength of the interaction which has to be specified.

In the Bayesian paradigm, one principled approach to handle hyperparameters is hierarchical modeling, by assuming a hyperprior and treating them as a part of the inference procedure. Specifically, we write the Gaussian prior $p(\mathbf{x}|\alpha) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \alpha^{-1}\bar{\mathbf{C}}_0)$, and employ a Gamma distribution $p(\alpha|a, b) = \mathrm{Gamma}(\alpha|a, b)$ on α, where (a, b) are the parameters. The Gamma distribution is the conjugate prior for α, and it is analytically and computationally convenient. In practice, one may take (a, b) close to $(1, 0)$ to mimic a noninformative prior. Then appealing to Bayes' formula again, one obtains a posterior distribution (jointly over $(\mathbf{x}, \alpha)$). Conceptually, with the VGA, this construction determines the optimal parameter by maximizing ELBO as a function of α, i.e. model selection within a parametric family. Thus it can be viewed as a direct application of ELBO in model selection.

One may explore the resulting joint posterior distribution in several ways [19, chapter 7]. In this work, we employ an EM type method to maximize the following (joint) lower bound

Algorithm 2. Hierarchical variational Gaussian approximation.

1: Input $(\mathbf{A}, \mathbf{y})$, and initialize α^1
2: **for** $k = 1, 2, \ldots$ **do**
3: E-step: Update $(\bar{\mathbf{x}}^k, \mathbf{C}^k)$ by algorithm 1:
$$(\bar{\mathbf{x}}^k, \mathbf{C}^k) = \arg \max_{(\bar{\mathbf{x}}, \mathbf{C}) \in \mathbb{R}^m \times \mathcal{S}_m^+} F_{\alpha^k}(\bar{\mathbf{x}}, \mathbf{C});$$
4: M-step: Update α by (5.2);
5: Check the stopping criterion.
6: **end for**
7: Output: $(\bar{\mathbf{x}}, \mathbf{C})$

$$\begin{aligned} F(\bar{\mathbf{x}}, \mathbf{C}, \alpha) &= \int q(\mathbf{x}) \ln \frac{p(\mathbf{x}, \mathbf{y}|\alpha)\, p(\alpha|a, b)}{q(\mathbf{x})} \mathrm{d}\mathbf{x} \\ &= \int q(\mathbf{x}) \ln \frac{p(\mathbf{x}, \mathbf{y}|\alpha)}{q(\mathbf{x})} \mathrm{d}\mathbf{x} + \int q(\mathbf{x}) \ln p(\alpha|a, b) \mathrm{d}\mathbf{x} \\ &= F_\alpha(\bar{\mathbf{x}}, \mathbf{C}) + (a-1)\ln\alpha - \alpha b + \ln \frac{b^a}{\Gamma(a)}, \end{aligned}$$

where the subscript α indicates the dependence of ELBO on α. Then, using (3.4) and substituting $\mathbf{C}_0$ with $\alpha^{-1}\bar{\mathbf{C}}_0$, we have

$$\begin{aligned} F(\bar{\mathbf{x}}, \mathbf{C}, \alpha) = (\mathbf{y}, \mathbf{A}\bar{\mathbf{x}}) - (\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)}) - \frac{\alpha}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^t \bar{\mathbf{C}}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) - \frac{\alpha}{2}\mathrm{tr}(\bar{\mathbf{C}}_0^{-1}\mathbf{C}) \\ + \frac{1}{2}\ln|\mathbf{C}| + \frac{m}{2}\ln\alpha - \frac{1}{2}\ln|\bar{\mathbf{C}}_0| + (a-1)\ln\alpha - \alpha b + \frac{m}{2} - (\mathbf{1}_n, \ln(\mathbf{y}!)) + \ln\frac{b^a}{\Gamma(a)}. \end{aligned} \tag{5.1}$$

This functional extends ELBO $F(\bar{\mathbf{x}}, \mathbf{C})$ to estimate the hyperparameter α simultaneously with $(\bar{\mathbf{x}}, \mathbf{C})$ in a way analogous to augmented Tikhonov regularization [22].

To maximize $F(\bar{\mathbf{x}}, \mathbf{C}, \alpha)$, we employ an EM algorithm [6, chapter 9.3]. In the E-step, we fix α, and maximize $F(\bar{\mathbf{x}}, \mathbf{C}, \alpha)$ for a new Gaussian approximation $\mathcal{N}(\mathbf{x}|\bar{\mathbf{x}}, \mathbf{C})$ by algorithm 1, with the unique maximizer denoted by $(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$. Then in the M-step, we fix $(\bar{\mathbf{x}}, \mathbf{C})$ and update α by

$$\alpha = \frac{m + 2(a-1)}{(\bar{\mathbf{x}}_\alpha - \boldsymbol{\mu}_0)^t \bar{\mathbf{C}}_0^{-1}(\bar{\mathbf{x}}_\alpha - \boldsymbol{\mu}_0) + \mathrm{tr}(\bar{\mathbf{C}}_0^{-1}\mathbf{C}_\alpha) + 2b}. \tag{5.2}$$

This follows from the condition $\frac{\partial F}{\partial \alpha} = 0$. These discussions lead to the procedure in algorithm 2. A natural stopping criterion at line 5 is the change of α. Below we analyze the convergence of algorithm 2.

Remark 5.1. The first two terms in the denominator of the iteration (5.2) is given by

$$\alpha(\bar{\mathbf{x}}_\alpha - \boldsymbol{\mu}_0)^t \bar{\mathbf{C}}_0^{-1}(\bar{\mathbf{x}}_\alpha - \boldsymbol{\mu}_0) + \alpha\mathrm{tr}(\bar{\mathbf{C}}_0^{-1}\mathbf{C}_\alpha) = \mathbb{E}_{q(\mathbf{x})}[\|\mathbf{x} - \boldsymbol{\mu}_0\|^2_{\mathbf{C}_0^{-1}}],$$

i.e. the expectation of the negative logarithm of the Gaussian prior $p(\mathbf{x})$ with respect to the Gaussian posterior approximation $q(\mathbf{x})$. Formally, the fixed point iteration (5.2) can be viewed as an extension of that for a balancing principle for Tikhonov regularization in [20, 22] to a probabilistic context.

In order to analyze the convergence of algorithm 2, we write the functional $F_\alpha(\bar{\mathbf{x}}, \mathbf{C})$ as

$$F_\alpha(\bar{\mathbf{x}}, \mathbf{C}) = \phi(\bar{\mathbf{x}}, \mathbf{C}) + \alpha\psi(\bar{\mathbf{x}}, \mathbf{C}),$$

where

$$\phi(\bar{\mathbf{x}}, \mathbf{C}) = (\mathbf{y}, \mathbf{A}\bar{\mathbf{x}}) - (\mathbf{1}_n, \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)}) + \frac{1}{2}\ln|\mathbf{C}| - \frac{1}{2}\ln|\bar{\mathbf{C}}_0| - (\mathbf{1}_n, \ln(\mathbf{y}!)),$$
$$\psi(\bar{\mathbf{x}}, \mathbf{C}) = -\frac{1}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^t\bar{\mathbf{C}}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) - \frac{1}{2}\mathrm{tr}(\bar{\mathbf{C}}_0^{-1}\mathbf{C}) \leqslant 0.$$

Thus the functional $F_\alpha(\bar{\mathbf{x}}, \mathbf{C})$ resembles classical Tikhonov regularization. By theorem 3.2, for any $\alpha > 0$, there exists a unique maximizer $(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ to F_α, and the value function $\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ is continuous in α, see lemma 5.2 below. In appendix B, we show that the maximizer $(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ is actually differentiable in α.

Lemma 5.1. *For any $\alpha > 0$, the maximizer $(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ is bounded, with the bound depending only on α.*

Proof. Taking inner product between (4.1) and $\bar{\mathbf{x}}_\alpha$, we deduce

$$(\mathbf{C}_0^{-1}\bar{\mathbf{x}}_\alpha, \bar{\mathbf{x}}_\alpha) + (\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}_\alpha + \mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)}, \mathbf{A}\bar{\mathbf{x}}_\alpha) = (\mathbf{A}^t\mathbf{y}, \bar{\mathbf{x}}_\alpha).$$

It can be verified directly that the function $f(t) = te^t$ is bounded from below by $-e^{-1}$ for $t \in \mathbb{R}$. Meanwhile, by (4.2), $\mathbf{C} \leqslant \mathbf{C}_0$, and thus

$$(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}}_\alpha + \mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)}, \mathbf{A}\bar{\mathbf{x}}_\alpha) \geqslant -e^{-1}\sum_i \mathrm{e}^{\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)_i} \geqslant -e^{-1}\sum_i \mathrm{e}^{\mathrm{diag}(\mathbf{A}\mathbf{C}_0\mathbf{A}^t)_i} = -ce^{-1}.$$

This and the Cauchy–Schwarz inequality give $\|\bar{\mathbf{x}}_\alpha\| \leqslant c\alpha^{-1}$, with c depending only on $\mathbf{y}$. Next, by (4.2), we have

$$0 \leqslant \mathrm{e}^{(\mathbf{A}\bar{\mathbf{x}})_i + \mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)_i} \leqslant \mathrm{e}^{(\mathbf{A}\bar{\mathbf{x}})_i + \mathrm{diag}(\mathbf{A}\mathbf{C}_0\mathbf{A}^t)_i} \leqslant c,$$

and consequently appealing to (4.2) again yields $(\mathbf{C}_0^{-1} + c\mathbf{A}^t\mathbf{A})^{-1} \leqslant \mathbf{C} \leqslant \mathbf{C}_0$, completing the proof. □

Lemma 5.2. *The functional value $\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ is continuous at any $\alpha > 0$.*

Proof. Let $\{\alpha^k\} \subset \mathbb{R}^+$ be a sequence convergent to α. By theorem 3.2, for each α^k, there exists a unique maximizer $(\bar{\mathbf{x}}^k, \mathbf{C}^k)$ to $F_{\alpha^k}(\bar{\mathbf{x}}, \mathbf{C})$. By lemma 5.1, the sequence $\{(\bar{\mathbf{x}}^k, \mathbf{C}^k)\}$ is uniformly bounded, and there exists a convergent subsequence, relabeled as $\{(\bar{\mathbf{x}}^k, \mathbf{C}^k)\}$, with a limit $(\bar{\mathbf{x}}^*, \mathbf{C}^*)$. By the continuity of the functionals $\phi(\bar{\mathbf{x}}, \mathbf{C})$ and $\psi(\bar{\mathbf{x}}, \mathbf{C})$, we have for any $(\bar{\mathbf{x}}, \mathbf{C}) \in \mathbb{R}^m \times \mathcal{S}_m^+$

$$\begin{aligned} F_\alpha(\bar{\mathbf{x}}^*, \mathbf{C}^*) &= \lim_{k\to\infty}(\phi(\bar{\mathbf{x}}^k, \mathbf{C}^k) + \alpha_k\psi(\bar{\mathbf{x}}^k, \mathbf{C}^k)) \geqslant \lim_{k\to\infty}(\phi(\bar{\mathbf{x}}, \mathbf{C}) + \alpha_k\psi(\bar{\mathbf{x}}, \mathbf{C})) \\ &= \phi(\bar{\mathbf{x}}, \mathbf{C}) + \alpha\psi(\bar{\mathbf{x}}, \mathbf{C}) = F_\alpha(\bar{\mathbf{x}}, \mathbf{C}). \end{aligned}$$

That is, $(\bar{\mathbf{x}}^*, \mathbf{C}^*)$ is a maximizer of $F_\alpha(\bar{\mathbf{x}}, \mathbf{C})$. The uniqueness of the maximizer to $F_\alpha(\bar{\mathbf{x}}, \mathbf{C})$ and a standard subsequence argument imply that the whole sequence converges. The desired continuity now follows by the continuity of $\psi(\bar{\mathbf{x}}, \mathbf{C})$ in $(\bar{\mathbf{x}}, \mathbf{C})$. □

Next we give an important monotonicity relation for $\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ in α, in a manner similar to classical Tikhonov regularization [20]. In appendix B, we show that it is actually strictly monotone.

Lemma 5.3. *The functional $\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ is monotonically increasing in α.*

Proof. This result follows by a standard comparison principle. For any α_1, α_2, by the maximizing property of $(\mathbf{C}_{\alpha_1}, \bar{\mathbf{x}}_{\alpha_1})$ and $(\mathbf{C}_{\alpha_2}, \bar{\mathbf{x}}_{\alpha_2})$, we have

$$F_{\alpha_1}(\bar{\mathbf{x}}_{\alpha_1}, \mathbf{C}_{\alpha_1}) \geqslant F_{\alpha_1}(\bar{\mathbf{x}}_{\alpha_2}, \mathbf{C}_{\alpha_2}) \quad \text{and} \quad F_{\alpha_2}(\bar{\mathbf{x}}_{\alpha_2}, \mathbf{C}_{\alpha_2}) \geqslant F_{\alpha_2}(\bar{\mathbf{x}}_{\alpha_1}, \mathbf{C}_{\alpha_1}).$$

Summing up these two inequalities and collecting terms yield

$$(\alpha_1 - \alpha_2)[\psi(\bar{\mathbf{x}}_{\alpha_1}, \mathbf{C}_{\alpha_1}) - \psi(\bar{\mathbf{x}}_{\alpha_2}, \mathbf{C}_{\alpha_2})] \geqslant 0.$$

Then the desired monotonicity relation follows. □

Theorem 5.1. *For any initial guess $\alpha^1 > 0$, the sequence $\{\alpha^k\}$ generated by algorithm 2 is monotonically convergent to some $\alpha^* \geqslant 0$, and if the limit $\alpha^* > 0$, then it satisfies the fixed point equation (5.2).*

Proof. By the fixed point iteration (5.2), we have (with $c = \frac{m}{2} + a - 1$)

$$\begin{aligned}\alpha^{k+1} - \alpha^k &= \frac{c}{-\psi(\bar{\mathbf{x}}_{\alpha^k}, \mathbf{C}_{\alpha^k}) + b} - \frac{c}{-\psi(\bar{\mathbf{x}}_{\alpha^{k-1}}, \mathbf{C}_{\alpha^{k-1}}) + b} \\ &= \frac{c[\psi(\bar{\mathbf{x}}_{\alpha^k}, \mathbf{C}_{\alpha^k}) - \psi(\bar{\mathbf{x}}_{\alpha^{k-1}}, \mathbf{C}_{\alpha^{k-1}})]}{(-\psi(\bar{\mathbf{x}}_{\alpha^k}, \mathbf{C}_{\alpha^k}) + b)(-\psi(\bar{\mathbf{x}}_{\alpha^{k-1}}, \mathbf{C}_{\alpha^{k-1}}) + b)}.\end{aligned}$$

Since $\psi \leqslant 0$, the denominator is positive. By lemma 5.3, $\alpha^{k+1} - \alpha^k$ and $\alpha^k - \alpha^{k-1}$ have the same sign, and thus $\{\alpha^k\}$ is monotone. Further, for all α^k, we have $0 \leqslant \alpha^k \leqslant \frac{m+2(a-1)}{2b}$, i.e. $\{\alpha^k\}$ is uniformly bounded. Thus $\{\alpha^k\}$ is convergent. By lemma 5.2, $\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ is continuous in α for $\alpha > 0$, and α^* satisfies (5.2). □

Remark 5.2. The proof of theorem 5.1 provides a constructive approach to the existence of a solution to (5.2). The uniqueness of the solution α^* to (5.2) is generally not ensured. However, in practice, it seems to have only two fixed points: one is in the neighborhood of $+\infty$, which is uninteresting, and the other is the desired one.

6. Numerical experiments and discussions

Now we present numerical results to examine algorithmic features (sections 6.1–6.4, with the example `phillips`) and to illustrate the VGA (section 6.5). All one-dimensional examples are taken from public domain `MATLAB` package `Regutools`[3], and the discrete problems are of size 100×100. We refer the prior with a zero mean $\boldsymbol{\mu}_0 = \mathbf{0}$ and the covariance $\alpha^{-1}\mathbf{I}_m$ and $\alpha^{-1}\mathbf{L}_1^{-1}\mathbf{L}_1^{-t}$ (with $\mathbf{L}_1$ being the 1D first-order forward difference matrix) to as the L^2- and H^1-prior, respectively, and let $\bar{\mathbf{C}}_0 = \mathbf{I}_m$, and $\bar{\boldsymbol{C}}_1 = \mathbf{L}_1^{-1}\mathbf{L}_1^{-t}$. Unless otherwise specified, the parameter α is determined in a trial-and-error manner, and in algorithm 1, the Newton update

[3] www.imm.dtu.dk/~pcha/Regutools/, last accessed on April 15, 2017.

$\delta\bar{\mathbf{x}}$ in (4.3) is computed by the `MATLAB` built-in function `pcg` with a default tolerance, the prior covariance $\mathbf{C}_0^{-1}$ as the preconditioner and a maximum 10 PCG iterations.

6.1. Convergence behavior of inner and outer iterations of algorithm 1

First, we examine the convergence behavior of inner iterations for updating $\bar{\mathbf{x}}$ and $\mathbf{C}$, i.e. (4.3) and (4.4), for the example `phillips` with the L^2-prior $\mathbf{C}_0 = 1.0 \times 10^{-1}\bar{\mathbf{C}}_0$ and H^1-prior $\mathbf{C}_0 = 2.5 \times 10^{-3}\bar{\mathbf{C}}_1$. To study the convergence, we fix $\mathbf{C}$ at $\mathbf{C}^1 = \mathbf{I}$ for $\bar{\mathbf{x}}$ and present the ℓ^2-norm of the update $\delta\bar{\mathbf{x}}$ (initialized with $\bar{\mathbf{x}}^0 = \mathbf{0}$), and similarly fix $\bar{\mathbf{x}}$ at the converged iterate $\bar{\mathbf{x}}^1$ for $\mathbf{C}$ and present the spectral norm of the change $\delta\mathbf{C}$. For both (4.3) and (4.4), these initial guesses are quite far away from the solutions, and thus the choice allows showing their global convergence behavior. The convergence is fairly rapid and steady for both inner iterations, see figure 1. For example, for a tolerance 10^{-5}, the Newton method (4.3) converges after about 10 iterations, and the fixed point method (4.4) converges after 4 iterations, respectively. The global as well as local superlinear convergence of the Newton method (4.3) are clearly observed, confirming the discussions in section 4. The convergence behavior of the inner iterations is similar for both priors. In practice, it is unnecessary to solve the inner iterates to a very high accuracy, and it suffices to apply a few inner updates within each outer iteration. Since the iteration (4.4) often converges faster than (4.1), we take five Newton updates and one fixed point update per outer iteration for the numerical experiments below.

To examine the convergence of outer iterations, we show the errors of the mean $\bar{\mathbf{x}}$ and covariance $\mathbf{C}$ and the lower bound $F(\bar{\mathbf{x}}, \mathbf{C})$ in figures 2 and 3, respectively. Algorithm 1 is terminated when the change of the lower bound falls below 10^{-10}. For the L^2-prior, algorithm 1 converges after 5 iterations and the last increments $\delta\bar{\mathbf{x}}$ and $\delta\mathbf{C}$ are of order 10^{-8} and 10^{-9}, respectively. This observation holds also for the H^1-prior, see figures 2(b) and 3(b). Thus, both inner and outer iterations converge rapidly and steadily, and algorithm 1 is very efficient.

6.2. Low-rank approximation of $\mathbf{A}$ and sparsity of $\mathbf{C}$

The discussions in section 4.2 show that the structure on $\mathbf{A}$ and $\mathbf{C}$ can be leveraged to reduce the complexity of algorithm 1. Now we evaluate their influence on the accuracy of the VGA.

First, we examine the influence of low-rank approximation to $\mathbf{A}$. Since the kernel function of the example `phillips` is smooth, the inverse problem is mildly ill-posed and the singular values σ_k decay algebraically, see figure 4(a). A low-rank matrix $\mathbf{A}_r$ of rank $r \approx 10$ can already approximate $\mathbf{A}$ well. To study its influence on the VGA, we denote by $(\bar{\mathbf{x}}_r, \mathbf{C}_r)$ and $(\bar{\mathbf{x}}^*, \mathbf{C}^*)$ the VGA for $\mathbf{A}_r$ and $\mathbf{A}$, respectively. The errors $e_{\bar{\mathbf{x}}} = \|\bar{\mathbf{x}}_r - \bar{\mathbf{x}}^*\|$ and $e_{\mathbf{C}} = \|\mathbf{C}_r - \mathbf{C}^*\|$ for different ranks r are shown in figures 4(b) and (c) for the L^2- and H^1-prior, respectively. Too small a rank r of the approximation $\mathbf{A}_r$ can lead to pronounced errors in both the mean $\bar{\mathbf{x}}$ and the covariance $\mathbf{C}$, whereas for a rank of $r = 10$, the errors already fall below one percent. Interestingly, the decay of the error $e_{\bar{\mathbf{x}}}$ is much faster than that of the singular values σ_k, and the error $e_{\mathbf{C}}$ decays slower than $e_{\bar{\mathbf{x}}}$. The fast decay of the errors $e_{\bar{\mathbf{x}}}$ and $e_{\mathbf{C}}$ indicates the robustness of the VGA, which justifies using low-rank approximations in algorithm 1.

Next we examine the influence of the sparsity assumption on the covariance $\mathbf{C}$, which is used to reduce the complexity of algorithm 1. Due to the coupling between $\bar{\mathbf{x}}$ and $\mathbf{C}$, see (4.1) and (4.2), the sparsity assumption on $\mathbf{C}$ affects the accuracy of both $\bar{\mathbf{x}}$ and $\mathbf{C}$. To illustrate this, we take different sparsity levels s on $\mathbf{C}$ in algorithm 1, i.e. at most s nonzero entries around the

Table 1. The errors $e_{\bar{\mathbf{x}}}$ and $e_{\mathbf{C}}$ versus the sparsity level s of $\mathbf{C}$ for `phillips`.

Prior	L^2 prior		H^1 prior	
s	$e_{\bar{\mathbf{x}}}$	$e_{\mathbf{C}}$	$e_{\bar{\mathbf{x}}}$	$e_{\mathbf{C}}$
1	6.38×10^{-2}	9.20×10^{-2}	1.92×10^{-2}	7.06×10^{-2}
3	5.62×10^{-2}	8.10×10^{-2}	1.27×10^{-2}	5.42×10^{-2}
5	4.88×10^{-2}	7.02×10^{-2}	1.00×10^{-2}	4.29×10^{-2}

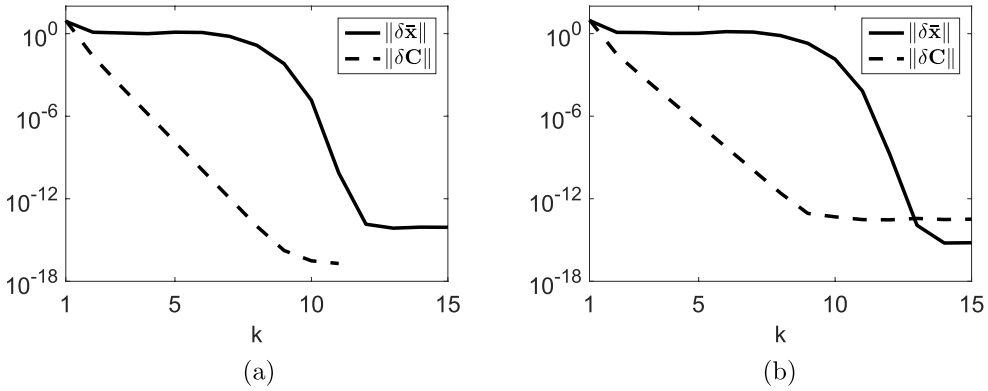

Figure 1. The convergence of the inner iterations of algorithm 1 for `phillips`. (a) L^2-prior and (b) H^1-prior.

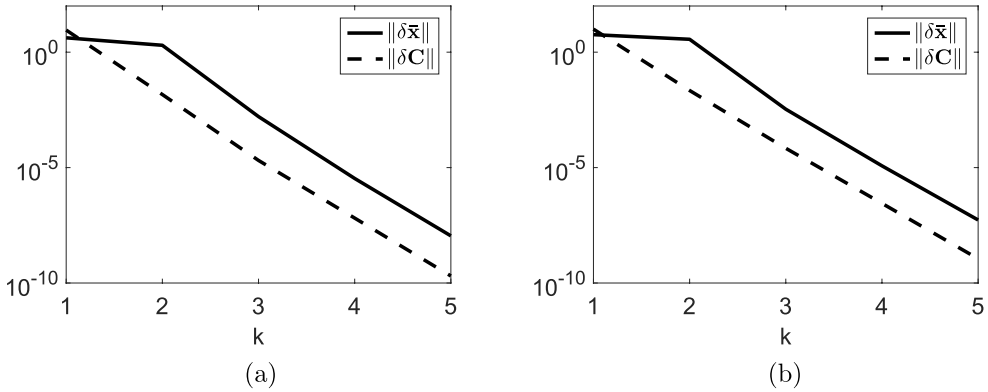

Figure 2. The convergence of outer iterations of algorithm 1 for `phillips`. (a) L^2 prior and (b) H^1-prior.

diagonal of $\mathbf{C}$. Surprisingly, a diagonal $\mathbf{C}$ already gives an acceptable approximation measured by the errors $e_{\bar{\mathbf{x}}} = \|\bar{\mathbf{x}}_s - \bar{\mathbf{x}}^*\|_2$ and $e_{\mathbf{C}} = \|\mathbf{C}_s - \mathbf{C}^*\|_2$, where $(\bar{\mathbf{x}}_s, \mathbf{C}_s)$ is the VGA with a sparsity level s. The errors $e_{\bar{\mathbf{x}}}$ and $e_{\mathbf{C}}$ decrease with the sparsity level s, see table 1. Thus the sparsity assumption on $\mathbf{C}$ can reduce significantly the complexity while retaining the accuracy.

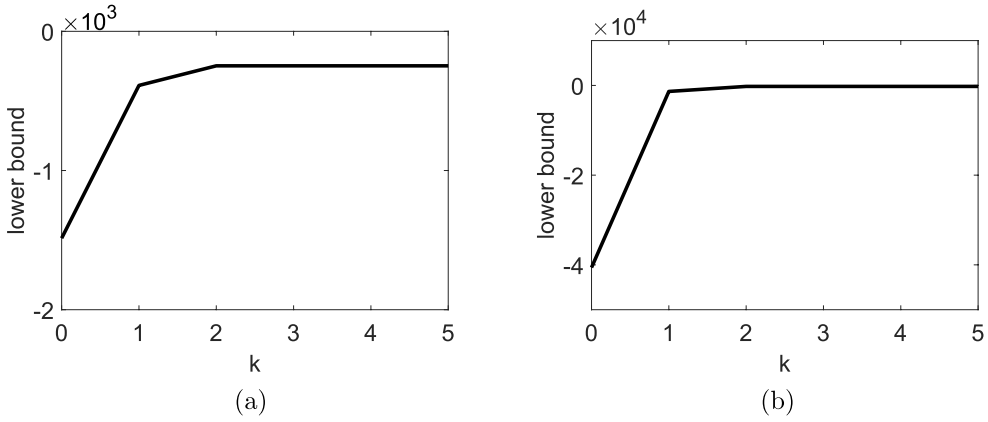

Figure 3. The convergence of the lower bound $F(\bar{\mathbf{x}}, \mathbf{C})$ for `phillips`. (a) L^2-prior and (b) H^1-prior.

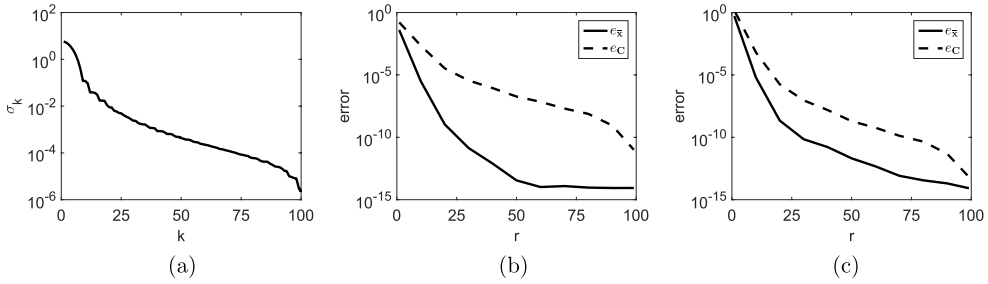

Figure 4. (a) singular values and (b)–(c): the errors of the mean and covariance for `phillips`. (a) singular values σ_k, (b) L^2-prior and (c) H^1-prior.

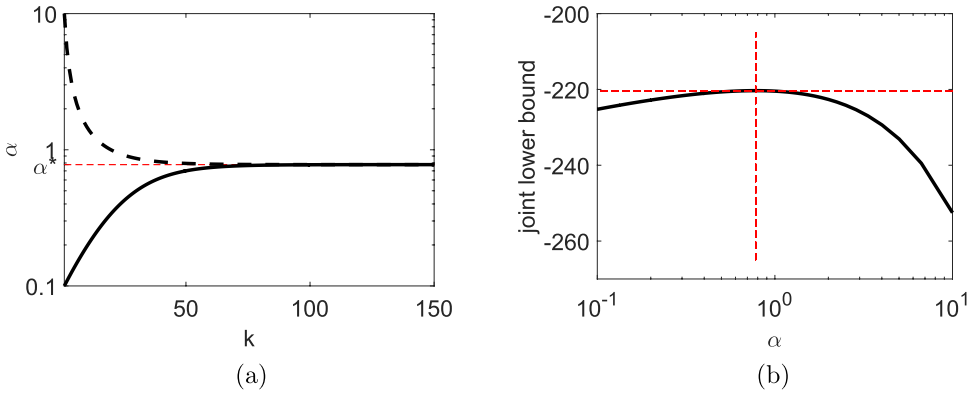

Figure 5. (a) The convergence of algorithm 2 initialized with 0.1 and 10, both convergent to $\alpha^* = 0.7778$ (b) the joint lower bound versus α, for `phillips` with L^2-prior. (a) convergence of α and (b) joint lower bound.

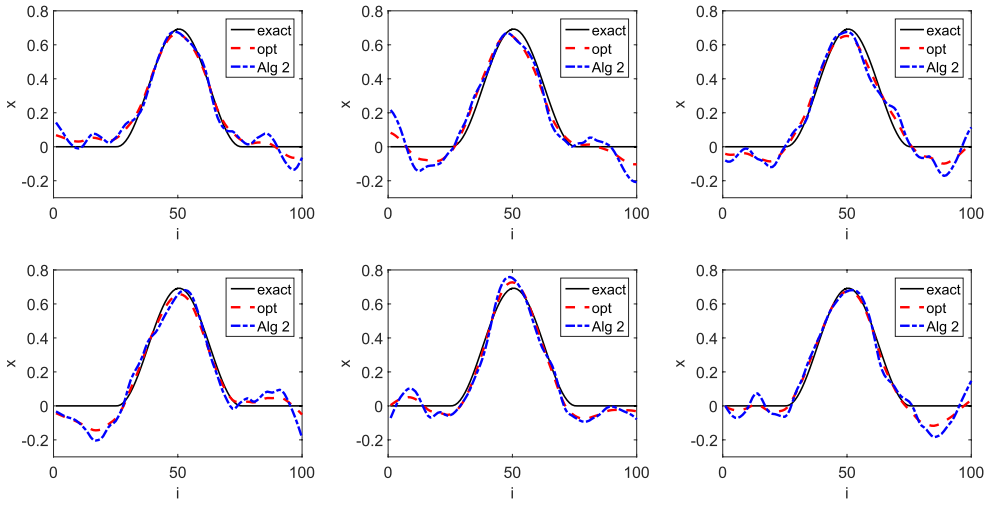

Figure 6. The mean $\bar{\mathbf{x}}$ of the Gaussian approximation by algorithm 2 (Alg2) and the 'optimal' solution (opt) for 6 realizations of Poisson data for `phillips` with the L^2-prior.

Table 2. The values of the hyperparameter α for the results in figure 6.

Case	1	2	3	4	5	6
Opt	2.64	3.35	2.59	1.35	9.31	4.04
Alg 2	0.78	0.76	0.76	0.77	0.73	0.74

6.3. Hierarchical parameter choice

Now we examine the convergence of algorithm 2 for choosing the parameter α in the prior $p(\mathbf{x})$. By theorem 5.1, the sequence $\{\alpha^k\}$ generated by algorithm 2 is monotone. We illustrate this by two initial guesses, i.e. $\alpha^1 = 0.1$ and $\alpha^1 = 10$. Both sequences of iterates generated by algorithm 2 converge monotonically to the limit $\alpha^* = 0.7778$, and the convergence of algorithm 2 is fairly steady, see figure 5(a). Further, algorithm 2 indeed maximizes the joint lower bound (5.1) with its maximum attained at $\alpha^* = 0.7778$, see figure 5(b). Though not shown, the lower bound $F_\alpha(\bar{\mathbf{x}}, \mathbf{c}|\alpha)$ is also increasing during the iteration. Thus, the hierarchical approach is indeed performing model selection by maximizing ELBO.

To illustrate the quality of the automatically chosen parameter α, we take six realizations of the Poisson data $\mathbf{y}$ and compare the mean $\bar{\mathbf{x}}$ of the VGA with the optimal regularized solutions, where α is tuned so that the error is smallest (and thus it is infeasible in practice). The means $\bar{\mathbf{x}}$ by algorithm 2 are comparable with the optimal ones, see figure 6, and thus the hierarchical approach can yield reasonable approximations. The parameter α by the hierarchical approach is slightly smaller than the optimal one, see table 2, and hence the corresponding reconstruction tends to be slightly more oscillatory than the optimal one. The value of the parameter α by the hierarchical approach is relatively independent of the realization, whose precise mechanism is to be ascertained.

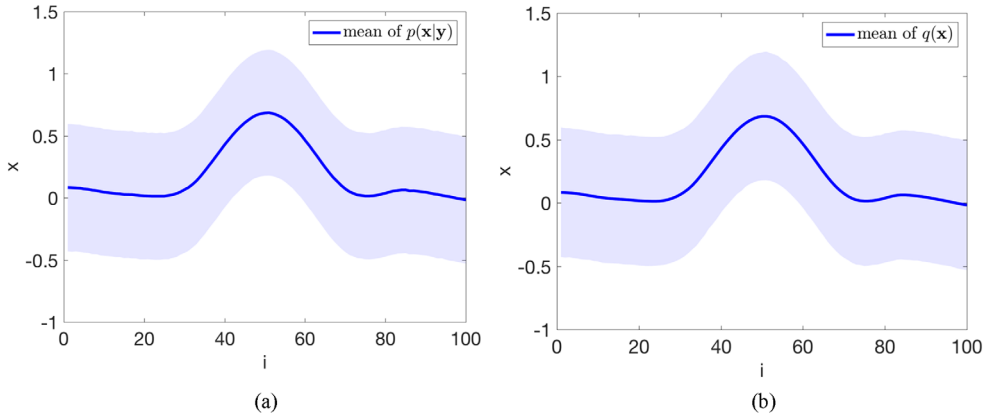

Figure 7. The mean and 90% HPD credible set by (a) MCMC and (b) VGA for `phillips` with $\mathbf{C}_0 = 1.00 \times 10^{-1}\bar{\mathbf{C}}_0$.

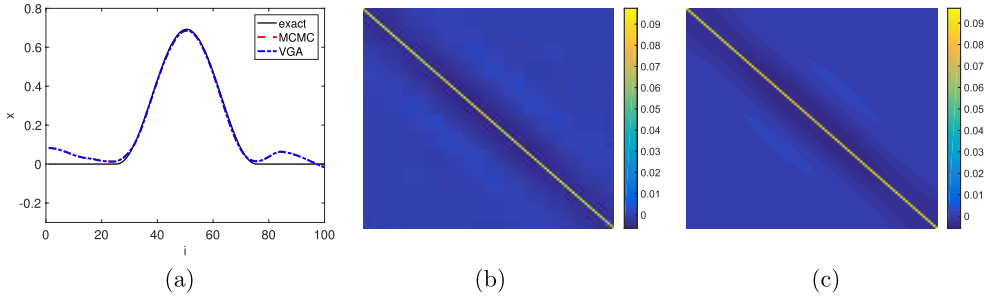

Figure 8. (a) The mean by MCMC and VGA versus the exact solution, and the covariance by (b) MCMC and (c) VGA for `phillips` with $\mathbf{C}_0 = 1.00 \times 10^{-1}\bar{\mathbf{C}}_0$.

6.4. VGA versus MCMC

Despite the widespread use of variational type techniques in practice, the accuracy of the approximations is rarely theoretically studied. This has long been a challenging issue for approximate Bayesian inference, including the VGA. In this part, we conduct an experiment to numerically validate the VGA against the results by Markov chain Monte Carlo (MCMC). To this end, we employ the standard Metropolis–Hastings algorithm, with the Gaussian approximation from the VGA as the proposal distribution (i.e. independence sampler). In other words, we correct the samples drawn from VGA by a Metropolis–Hastings step. The length of the MCMC chain is 2×10^5, and the last 1×10^5 samples are used for computing the summarizing statistics. The acceptance rate in the Metropolis–Hastings algorithm is 96.06%. This might be attributed to the fact that the VGA approximates the posterior distribution fairly accurately, and thus nearly all the proposals are accepted. The numerical results are presented in figure 7, where the mean and the 90% highest posterior density (HPD) credible set are shown, with the credible set computed componentwise. It is observed that the mean and HPD credible sets by MCMC and VGA are very close to each other, see figures 7 and 8, thereby validating the accuracy of the VGA. The ℓ^2 error between the mean by MCMC and GVA is 9.80×10^{-3}, and the error between corresponding covariance in spectral norm is 6.40×10^{-3}. Graphically the means and covariances are indistinguishable, see figure 8.

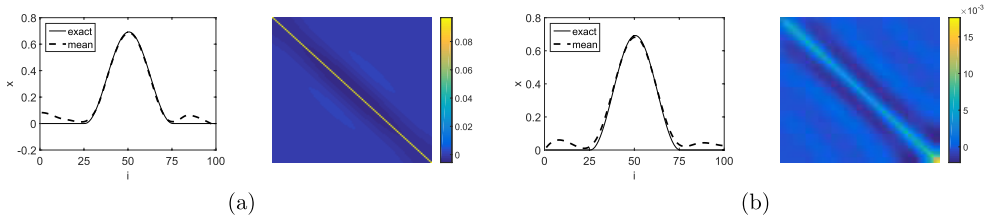

Figure 9. The Gaussian approximation for `phillips`. (a) $\mathbf{C}_0 = 1.00 \times 10^{-1}\bar{\mathbf{C}}_0$ and (b) $\mathbf{C}_0 = 2.5 \times 10^{-3}\bar{\mathbf{C}}_1$.

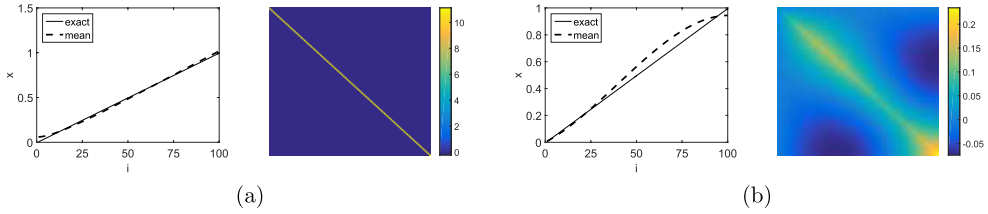

Figure 10. The Gaussian approximation for `foxgood`. (a) $\mathbf{C}_0 = 1.12 \times 10^{1}\bar{\mathbf{C}}_0$ and (b) $\mathbf{C}_0 = 9.8 \times 10^{-3}\bar{\mathbf{C}}_1$.

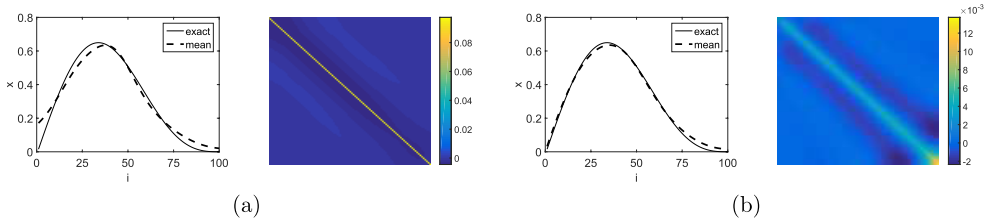

Figure 11. The Gaussian approximation for `gravity`. (a) $\mathbf{C}_0 = 1 \times 10^{-1}\bar{\mathbf{C}}_0$ and (b) $\mathbf{C}_0 = 1.5 \times 10^{-3}\bar{\mathbf{C}}_1$.

6.5. Numerical reconstructions

Last, we present VGAs for one- and two-dimensional examples. The numerical results for the following four 1D examples, i.e. `phillips`, `foxgood`, `gravity` and `heat`, for both L^2- and H^1-priors, are presented in figures 9–12. For the example `phillips` with either prior, the mean $\bar{\mathbf{x}}$ by algorithm 1 agrees very well with the true solution $\mathbf{x}^\dagger$. However, near the boundary, the mean $\bar{\mathbf{x}}$ is less accurate. This might be attributed to the fact that in these regions, the Poisson count is relatively small, and it may be insufficient for an accurate recovery. For the L^2-prior, the optimal $\mathbf{C}$ is diagonal dominant, and decays rapidly away from the diagonal, see figure 9(b). For the H^1-prior, $\mathbf{C}$ remains largely diagonally dominant, but the off-diagonal entries decay a bit slower. Thus, it is valid to assume that $\mathbf{C}$ is dominated by local interactions as in section 4.2. These observations remain largely valid for the other 1D examples, despite that they are much more ill-posed.

Last, we test algorithm 1 on a 2D image of size 128×128. In this example, the matrix $\mathbf{A} \in \mathbb{R}^{16\,384 \times 16\,384}$ is a (discrete) Gaussian blurring kernel with a blurring width 99, variance 1.5 and a circular boundary condition. Since the blurring width is large, the matrix $\mathbf{A}$ is indeed

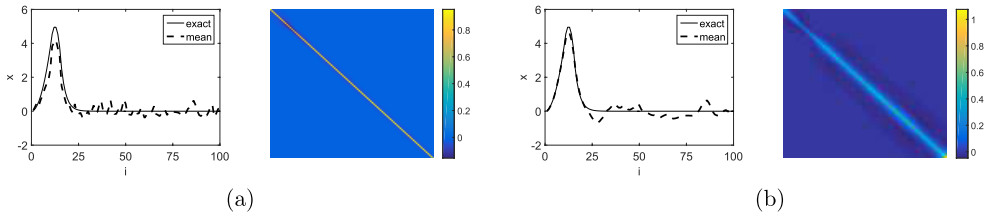

Figure 12. The Gaussian approximation for `heat`. (a) $\mathbf{C}_0 = 3.2 \times 10^{-1}\bar{\mathbf{C}}_0$ and (b) $\mathbf{C}_0 = 1 \times 10^{0}\bar{\mathbf{C}}_1$.

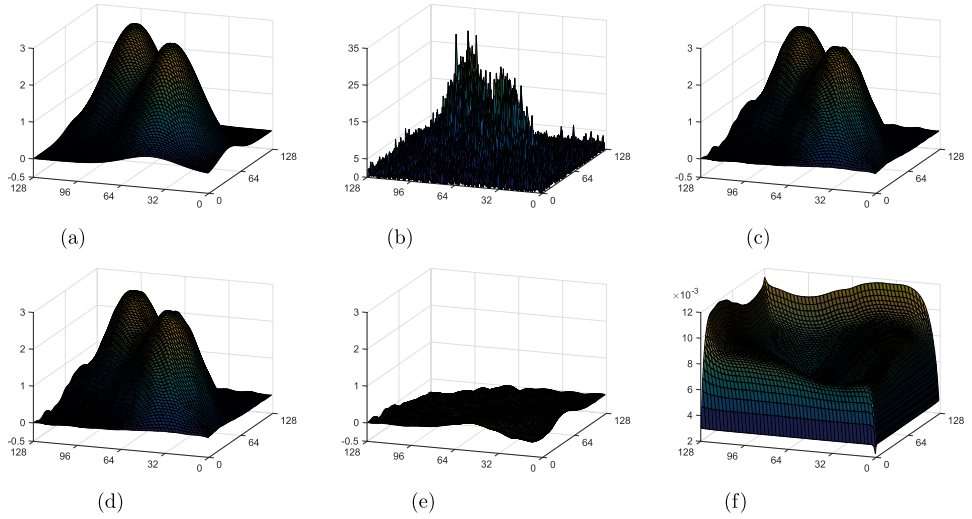

Figure 13. The Gaussian approximation for image deblurring. (a) True solution $\mathbf{x}^\dagger$, (b) Poisson sample $\mathbf{y}$, (c) MAP $\mathbf{x}_{\text{MAP}}$, (d) mean $\bar{\mathbf{x}}$, (e) error $\mathbf{x}^\dagger - \bar{\mathbf{x}}$ and (f) variance $\text{diag}(\mathbf{C})$.

low-rank, and we employ a rSVD approximation of rank 2000, where the rank is determined by inspecting the singular value spectrum. The true solution $\mathbf{x}^\dagger$ consists of two Gaussian blobs, see figure 13(a), and thus we employ a smooth prior with $\mathbf{C}_0 = 6.00 \times 10^{-2}\mathbf{L}^{-1}\mathbf{L}^{-t}$, where $\mathbf{L} = \mathbf{I} \otimes \mathbf{L}_1 + \mathbf{L}_1 \otimes \mathbf{I}$ is the 2D first-order finite difference matrix. Since the problem size is very large, we restrict $\mathbf{C}$ to be a sparse matrix such that every pixel interacts only with at most four neighboring pixels. This allows reducing the computational cost greatly. The mean $\bar{\mathbf{x}}$ is nearly identical with the true solution $\mathbf{x}^\dagger$, and the error is very small, see figure 13. We also compare the mean $\bar{\mathbf{x}}$ of the VGA solution with the MAP estimator $\hat{\mathbf{x}}$ in three different measures, i.e. ℓ_2 error, structural similarity index and PSNR, which are 9.72, 0.812 and 18.64 for $\bar{\mathbf{x}}$, respectively 9.74, 0.813, and 18.63 for $\hat{\mathbf{x}}$. These results indicate that the mean $\bar{\mathbf{x}}$ and the MAP estimator $\hat{\mathbf{x}}$ represent equally good approximations. To indicate the uncertainty around the mean $\bar{\mathbf{x}}$, we show in figure 13(f) the diagonal entries of $\mathbf{C}$ (i.e. the variance at each pixel). The variances are relatively large at pixels where the mean $\bar{\mathbf{x}}$ is less accurate.

In summary, the VGA can provide a reliable point estimator together with useful covariance estimates.

7. Conclusions

In this work, we have presented a study of the variational Gaussian approximation to the Poisson data (under the log linear link function) with respect to the Kullback–Leibler divergence. We derived explicit expressions for the lower bound functional and its gradient, and proved its strict concavity and existence and uniqueness of an optimal Gaussian approximation. Then we developed an efficient algorithm for maximizing the functional, discussed its convergence properties, and described practical strategies for reducing the complexity per iteration. Further, we analyzed hierarchical modeling for automatically determining the hyperparameter using the variational Gaussian approximation, and proposed a monotonically convergent algorithm for the joint estimation. The numerical experiments indicate that the algorithm converges rapidly, and the variational Gaussian approximation can accurately capture the posterior distribution.

There are several avenues for further study. First, one of fundamental issues is the quality of the Gaussian approximation relative to the true posterior distribution. In general this issue has been long standing, and it also remains to be analyzed for the Poisson model. Second, the variational Gaussian approximation can be viewed as a nonstandard regularization scheme, by also penalizing the covariance. This naturally motivates the study on its regularizing property from the perspective of classical regularization theory, e.g. consistency and convergence rates. Third, the approach generally gives a very reasonable approximation. This suggests itself as a preconditioner for sampling techniques, e.g. variational approximation as the proposal distribution (i.e. independence sampler) in the standard Metropolis–Hastings type algorithm or as the base distribution for importance sampler. It is expected to significantly speed up the convergence of these sampling procedures, which is confirmed by the preliminary experiments. We plan to study these aspects in future works.

Acknowledgments

The work of B Jin is supported by UK EPSRC grant EP/M025160/1, and that of C Zhang by a departmental studentship.

Appendix A. On the iteration (4.4)

In this appendix, we discuss an interesting property of the iteration (4.4), for the initial guess $\mathbf{C}^0 = \mathbf{C}_0$. We denote the fixed point map in (4.4) by $\mathbf{T}$, i.e.

$$\mathbf{T}(\mathbf{C}) = (\mathbf{C}_0^{-1} + \mathbf{A}^t \mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})\mathbf{A})^{-1}.$$

The next result gives the antimonotonicity of the map $\mathbf{T}$ on $\mathcal{S}_m^+$, i.e. for $\mathbf{C}, \tilde{\mathbf{C}} \in \mathcal{S}_m^+$, if $0 \leqslant \mathbf{C} \leqslant \tilde{\mathbf{C}}$, then $\mathbf{T}(\mathbf{C}) \geqslant \mathbf{T}(\tilde{\mathbf{C}})$.

Lemma A.1. *The mapping* $\mathbf{T}$ *is antimonotone.*

Proof. Let $\mathbf{C}, \tilde{\mathbf{C}} \in \mathcal{S}_m^+$. If $\mathbf{C} \leqslant \tilde{\mathbf{C}}$, then $\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t) \leqslant \mathrm{diag}(\mathbf{A}\tilde{\mathbf{C}}\mathbf{A}^t)$ componentwise. The claim follows from the identity $\mathbf{T}(\mathbf{C}) - \mathbf{T}(\tilde{\mathbf{C}}) = \mathbf{T}(\mathbf{C})\mathbf{A}^t \mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\tilde{\mathbf{C}}\mathbf{A}^t)} - \mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)})\mathbf{A}\mathbf{T}(\tilde{\mathbf{C}}) \geqslant 0$. □

The next result shows the monotonicity of the sequence $\{\mathbf{C}^k\}$ generated by (4.4).

Lemma A.2. *For any initial guess* $\mathbf{C}^0 \in \mathcal{S}_m^+$, *the sequence* $\{\mathbf{C}^k\}_{k\geqslant 0}$ *generated by the iteration (4.4) has the following properties: (i)* $\mathbf{C}^k \geqslant 0$ *for all* $k \geqslant 0$*; (ii)* $\mathbf{C}^k \leqslant \mathbf{C}_0$ *for all* $k \geqslant 0$*;*

(iii) If $\mathbf{C}^k \geqslant \mathbf{C}^j$ *then* $\mathbf{C}^{k+1} \leqslant \mathbf{C}^{j+1}$*; (iv) If* $\mathbf{C}^k \geqslant \mathbf{C}^j$ *then* $\mathbf{C}^{k+2} \geqslant \mathbf{C}^{j+2}$.

Proof. Properties (i) and (ii) are obvious. Properties (iii) and (iv) are direct consequences of the fact that the map $\mathbf{T}$ is antimonotone on $\mathcal{S}_m^+$, see lemma A.1. □

The next result shows that the sequence constitutes two subsequences, each converging to a fixed point of $\mathbf{T}^2$, which implies either a periodic orbit of period 2 of the map $\mathbf{T}$ or a fixed point of $\mathbf{T}$,

Theorem A.1. *With the initial guess* $\mathbf{C}^0 = \mathbf{C}_0$*, the sequence* $\{\mathbf{C}^k\}_{k\geqslant 0}$ *generated by iteration (4.4) converges to a fixed-point of* $\mathbf{T}^2$.

Proof. Lemma A.2(ii) implies

$$\mathbf{C}^2 \leqslant \mathbf{C}^0, \tag{A.1}$$

so we can use lemma A.2(iv) inductively to argue that $\{\mathbf{C}^{2k}\}_{k\geqslant 0}$ is a decreasing sequence. From (A.1) and lemma A.2(iii), we deduce $\mathbf{C}^1 \leqslant \mathbf{C}^3$, which together with lemma A.2(iv) implies that the sequence $\{\mathbf{C}^{2k+1}\}_{k\geqslant 0}$ is increasing. By the boundedness and monotonicity, both $\{\mathbf{C}^{2k}\}_{k\geqslant 0}$ and $\{\mathbf{C}^{2k+1}\}_{k\geqslant 0}$ converge, with the limit $\mathbf{C}^*$ and $\mathbf{C}^{**}$, respectively. These are the limits of the fixed point map $\mathbf{T}^2$. □

Remark A.1. By lemma A.2, $\mathbf{C}^* \geqslant \mathbf{C}^{**}$, and if $\mathbf{C}^* = \mathbf{C}^{**}$, the whole sequence converges. Generally, the interval of matrices $[\mathbf{C}^{**}, \mathbf{C}^*]$ provides a lower and sharp bounds for the fixed point of the iteration (4.4) (which is *a priori* known to be unique and to exist). By repeating the argument in [10, theorem 2.2], one may also examine the convergence of the sequence for the initial guess either $\mathbf{C}^0 < \mathbf{C}^{**}$ or $\mathbf{C}^0 > \mathbf{C}^*$.

Appendix B. Differentiability of the regularized solution

In this part, we discuss the differentiability of the regularized solution $(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ in α. For simplicity, we omit the subscript α. By differentiating (3.6) in α and chain rule, we obtain (with $\dot{\bar{\boldsymbol{x}}} = \frac{\mathrm{d}\bar{\mathbf{x}}}{\mathrm{d}\alpha}$ and $\dot{\mathbf{C}} = \frac{\mathrm{d}\mathbf{C}}{\mathrm{d}\alpha}$):

$$\begin{aligned}(\mathbf{A}^t\mathbf{D}\mathbf{A} + \alpha\mathbf{C}_0^{-1})\dot{\bar{\boldsymbol{x}}} + \frac{1}{2}\mathbf{A}^t\mathbf{D}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t) &= -\bar{\mathbf{C}}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0),\\ (\mathbf{C}^{-1}\dot{\mathbf{C}}\mathbf{C}^{-1} + \frac{1}{2}\mathbf{A}^t\mathbf{D}^{\frac{1}{2}}\mathrm{diagdiag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t)\mathbf{D}^{\frac{1}{2}}\mathbf{A}) + \mathbf{A}^t\mathbf{D}^{\frac{1}{2}}\mathrm{diag}(\mathbf{A}\dot{\bar{\mathbf{x}}})\mathbf{D}^{\frac{1}{2}}\mathbf{A} &= -\bar{\mathbf{C}}_0^{-1},\end{aligned} \tag{B.1}$$

where $\mathbf{D} = \mathrm{diag}(\mathrm{e}^{\mathbf{A}\bar{\mathbf{x}} + \frac{1}{2}\mathrm{diag}(\mathbf{A}\mathbf{C}\mathbf{A}^t)}) \in \mathbb{R}^{n\times n}$ is a diagonal matrix. This constitutes a coupled linear system for $(\dot{\bar{\boldsymbol{x}}}, \dot{\mathbf{C}})$. The next result gives its unique solvability.

Theorem B.1. *The sensitivity system (B.1) is uniquely solvable.*

Proof. Since the system (B.1) is linear and square, it suffices to show that the homogeneous problem has only a zero solution. To this end, by eliminating the variable $\dot{\bar{\boldsymbol{x}}}$ from the second line in (B.1) using the first line, we obtain the Schur complement for $\dot{\mathbf{C}}$:

$$\mathbf{C}^{-1}\dot{\mathbf{C}}\mathbf{C}^{-1} + \frac{1}{2}\mathbf{A}^t\mathbf{D}^{\frac{1}{2}}\mathrm{diagdiag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t)\mathbf{D}^{\frac{1}{2}}\mathbf{A} - \frac{1}{2}\mathbf{A}^t\mathbf{D}^{\frac{1}{2}}\mathrm{diag}(\mathbf{A}(\mathbf{A}^t\mathbf{D}\mathbf{A} + \alpha\bar{\mathbf{C}}_0^{-1})^{-1}\mathbf{A}^t\mathbf{D}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t))\mathbf{D}^{\frac{1}{2}}\mathbf{A}.$$

For any fixed $\mathbf{C}$, this defines a linear map on $\mathbb{R}^{m\times m}$. Next we show its invertibility. To this end, we take inner product the map with $\dot{\mathbf{C}}$, and show its positivity. Clearly, the first term is strictly positive. Thus it suffices to consider the last two terms. By the cyclic property of trace, with $\mathbf{d} = \mathrm{diag}(\mathbf{D}) \in \mathbb{R}^n$, we have

$$\begin{aligned}(\mathbf{A}^t\mathrm{diag}(\mathbf{d}\circ\mathrm{diag}(\mathbf{A}\dot{\boldsymbol{C}}\mathbf{A}^t))\mathbf{A}, \dot{\boldsymbol{C}}) &= \mathrm{tr}(\mathbf{A}^t\mathrm{diag}(\mathbf{d}\circ\mathrm{diag}(\mathbf{A}\dot{\boldsymbol{C}}\mathbf{A}^t))\mathbf{A}\dot{\mathbf{C}})\\ &= (\mathbf{D}\mathrm{diag}(\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t)), \mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t) = (\mathbf{D}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t), \mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t)) = (\bar{\mathbf{e}}, \bar{\mathbf{e}}),\end{aligned}$$

where $\bar{\mathbf{e}} = \mathbf{D}^{\frac{1}{2}}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t) \in \mathbb{R}^n$. Similarly, by letting $\bar{\mathbf{A}} = \mathbf{D}^{\frac{1}{2}}\mathbf{A}$, we have

$$\begin{aligned}&(\mathbf{A}^t\mathbf{D}\mathrm{diag}(\mathbf{A}(\mathbf{A}^t\mathbf{D}\mathbf{A} + \alpha\bar{\mathbf{C}}_0^{-1})^{-1}\mathbf{A}^t\mathbf{D}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\boldsymbol{A}^t))\mathbf{A}, \dot{\mathbf{C}})\\ &\quad= (\mathbf{D}\mathrm{diag}(\mathbf{A}(\mathbf{A}^t\mathbf{D}\mathbf{A} + \alpha\bar{\mathbf{C}}_0^{-1})^{-1}\mathbf{A}^t\mathbf{D}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t)), \mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t)\\ &\quad= (\bar{\mathbf{A}}(\bar{\mathbf{A}}^t\bar{\mathbf{A}} + \alpha\bar{\mathbf{C}}_0^{-1})^{-1}\bar{\mathbf{A}}^t\bar{\mathbf{e}}, \bar{\mathbf{e}}).\end{aligned}$$

Since $\mathbf{I}_n - \bar{\mathbf{A}}(\bar{\mathbf{A}}^t\bar{\mathbf{A}} + \alpha\bar{\mathbf{C}}_0^{-1})^{-1}\bar{\mathbf{A}}^t > 0$, the associated bilinear form is coercive on $\mathcal{S}_m^+$. Thus the Schur complement is invertible, and the system (B.1) has a unique solution. □

Corollary B.1. *For any rank deficient* $\mathbf{A}$, $\dot{\mathbf{C}} \neq \mathbf{0}$.

Proof. If $\dot{\mathbf{C}} = \mathbf{0}$, the second equation in (B.1) reduces to $\mathbf{A}^t\mathbf{D}^{\frac{1}{2}}\mathrm{diag}(\mathbf{A}\dot{\bar{\mathbf{x}}})\mathbf{D}^{\frac{1}{2}}\mathbf{A} = -\bar{\mathbf{C}}_0^{-1}$. By assumption, $\mathbf{A}$ is rank deficient, and thus the left hand side is rank deficient, whereas the right hand side is of full rank, which leads to a contradiction. Thus we have $\dot{\mathbf{C}} \neq \mathbf{0}$. □

The next result gives a lower-bound for the derivative $\frac{\mathrm{d}}{\mathrm{d}\alpha}\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$.

Theorem B.2. *The functional* $\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ *satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha) \geqslant \alpha(\mathbf{C}_0^{-1}\dot{\bar{\boldsymbol{x}}}, \dot{\bar{\boldsymbol{x}}}) + \frac{1}{2}(\mathbf{C}^{-1}\dot{\mathbf{C}}\mathbf{C}^{-1}, \dot{\mathbf{C}}).$$

Proof. By the definition of the functional ψ, we have

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha) = -(\bar{\mathbf{C}}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0), \dot{\bar{\boldsymbol{x}}}) - \frac{1}{2}(\bar{\mathbf{C}}_0^{-1}, \dot{\mathbf{C}}).$$

By taking inner product the first equation in (B.1) with $\dot{\bar{\mathbf{x}}}$, and the second with $\frac{1}{2}\dot{\mathbf{C}}$, we get

$$((\mathbf{A}^t\mathbf{D}\mathbf{A} + \alpha\mathbf{C}_0^{-1})\dot{\bar{\boldsymbol{x}}}, \dot{\bar{\boldsymbol{x}}}) + \frac{1}{2}(\mathbf{A}^t\mathbf{D}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t), \dot{\bar{\boldsymbol{x}}}) = -(\bar{\mathbf{C}}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0), \dot{\bar{\boldsymbol{x}}}),$$

$$\frac{1}{2}(\mathbf{C}^{-1}\dot{\mathbf{C}}\mathbf{C}^{-1} + \frac{1}{2}\mathbf{A}^t\mathbf{D}^{\frac{1}{2}}\mathrm{diagdiag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t)\mathbf{D}^{\frac{1}{2}}\mathbf{A}, \dot{\mathbf{C}}) + \frac{1}{2}(\mathbf{A}^t\mathbf{D}^{\frac{1}{2}}\mathrm{diag}(\mathbf{A}\dot{\bar{\boldsymbol{x}}})\mathbf{D}^{\frac{1}{2}}\mathbf{A}, \dot{\mathbf{C}}) = -\frac{1}{2}(\bar{\mathbf{C}}_0^{-1}, \dot{\mathbf{C}}).$$

By the cyclic property of trace and summing these two identities, we obtain

$$\begin{aligned}-(\bar{\mathbf{C}}_0^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0), \dot{\bar{\boldsymbol{x}}}) - \frac{1}{2}(\bar{\mathbf{C}}_0^{-1}, \dot{\mathbf{C}}) =& (\alpha\mathbf{C}_0^{-1}\dot{\bar{\boldsymbol{x}}}, \dot{\bar{\boldsymbol{x}}}) + \frac{1}{2}(\mathbf{C}^{-1}\dot{\mathbf{C}}\mathbf{C}^{-1}, \dot{\mathbf{C}})\\ &+ (\mathbf{A}^t\mathbf{D}\mathbf{A}\dot{\bar{\boldsymbol{x}}}, \dot{\bar{\boldsymbol{x}}}) + \frac{1}{4}(\mathbf{D}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t), \mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t))\\ &+ (\mathbf{D}^{\frac{1}{2}}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t), \mathbf{D}^{\frac{1}{2}}\mathbf{A}\dot{\bar{\boldsymbol{x}}}).\end{aligned} \tag{B.2}$$

Meanwhile, by the Cauchy–Schwarz inequality, we have

$$(\mathbf{D}^{\frac{1}{2}}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t), \mathbf{D}^{\frac{1}{2}}\mathbf{A}\dot{\bar{\mathbf{x}}}) \geqslant -(\mathbf{D}^{\frac{1}{2}}\mathbf{A}\dot{\bar{\mathbf{x}}}, \mathbf{D}^{\frac{1}{2}}\mathbf{A}\dot{\bar{\mathbf{x}}}) - \frac{1}{4}(\mathbf{D}^{\frac{1}{2}}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t), \mathbf{D}^{\frac{1}{2}}\mathrm{diag}(\mathbf{A}\dot{\mathbf{C}}\mathbf{A}^t)).$$

Substituting the preceding inequality into (B.2) yields the desired estimate. □

Corollary B.2. *The functional* $\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ *is strictly increasing in* α.

Proof. By theorem B.1, (B.1) is uniquely solvable. Since the right hand side of (B.1) is non-vanishing (by assumption, $\mathbf{C}_0$ is nonzero), the solution pair $(\dot{\bar{\mathbf{x}}}, \dot{\mathbf{C}})$ to (B.1) is nonzero. Thus, by theorem B.2, $\frac{\mathrm{d}}{\mathrm{d}\alpha}\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ is strictly positive, i.e. $\psi(\bar{\mathbf{x}}_\alpha, \mathbf{C}_\alpha)$ is strictly increasing. □

Remark B.1. For the standard regularized least-squares problem, the solution is distinct for different α, and it never vanishes (except the trivial case $\mathbf{y} = 0$). The proof in corollary B.2 indicates that an analogous statement holds for the Poisson model (2.2).

ORCID iDs

Bangti Jin https://orcid.org/0000-0002-3775-9155

References

[1] Anderson W N Jr, Kleindorfer G B, Kleindorfer P R and Woodroofe M B 1969 Consistent estimates of the parameters of a linear system *Ann. Math. Stat.* **40** 2064–75
[2] Archambeau C, Cornford D, Opper M and Shawe-Taylor J 2007 Gaussian process approximations of stochastic differential equations *JMLR: Workshop Conf. Proc.* vol 1 pp 1–16
[3] Barber D and Bishop C M 1998 Ensemble learning in Bayesian neural networks *NATO ASI Ser. F Comput. Syst. Sci.* **168** 215–38
[4] Bardsley J M and Luttman A 2016 A Metropolis–Hastings method for linear inverse problems with Poisson likelihood and Gaussian prior *Int. J. Uncertain. Quantif.* **6** 35–55
[5] Bertsekas D P 1998 *Nonlinear Programming* 2nd edn (Belmont, MA: Athena Scientific)
[6] Bishop C M 2006 *Pattern Recognition, Machine Learning* (Singapore: Springer)
[7] Cameron A and Trivedi P K 1998 *Regression Analysis of Count Data* (Cambridge: Cambridge University Press)
[8] Challis E and Barber D 2013 Gaussian Kullback–Leibler approximate inference *J. Mach. Learn. Res.* **14** 2239–86
[9] Dwyer P S 1967 Some applications of matrix derivatives in multivariate analysis *J. Am. Stat. Assoc.* **62** 607–25
[10] El-Sayed S M and Ran A C M 2001/02 On an iteration method for solving a class of nonlinear matrix equations *SIAM J. Matrix Anal. Appl.* **23** 632–45
[11] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Dordrecht: Kluwer)
[12] Erdoğan H and Fessler J A 1999 Monotonic algorithms for transmission tomography *IEEE Trans. Med. Imaging* **18** 801–14
[13] Gärtner B and Matousek J 2012 *Approximation Algorithms, Semidefinite Programming* (Berlin: Springer)
[14] Golub G H and Van Loan C F 2013 *Matrix Computations* 4th edn (Baltimore, MD: Johns Hopkins University Press)
[15] Halko N, Martinsson P G and Tropp J A 2011 Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions *SIAM Rev.* **53** 217–88

[16] Hall P, Ormerod J T and Wand M P 2011 Theory of Gaussian variational approximation for a Poisson mixed model *Stat. Sin.* **21** 369–89
[17] Hinton G E and Van Camp D 1993 Keeping the neural networks simple by minimizing the description length of the weights *Proc. 6th Annual Conf. Comput. Learning Theory* (ACM) (New York) pp 5–13
[18] Hohage T and Werner F 2016 Inverse problems with Poisson data: statistical regularization theory, applications and algorithms *Inverse Problems* **32** 093001
[19] Ito K and Jin B 2015 *Inverse Problems: Tikhonov Theory, Algorithms* (Hackensack, NJ: World Scientific)
[20] Ito K, Jin B and Takeuchi T 2011 A regularization parameter for nonsmooth Tikhonov regularization *SIAM J. Sci. Comput.* **33** 1415–38
[21] James W and Stein C 1961 Estimation with quadratic loss *Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability* vol I (Berkeley, CA: University California Press) pp 361–79
[22] Jin B and Zou J 2009 Augmented Tikhonov regularization *Inverse Problems* **25** 025001
[23] Kaipio J and Somersalo E 2005 *Statistical, Computational Inverse Problems* (New York: Springer)
[24] Kass R E and Raftery A E 1995 Bayes factors *J. Am. Stat. Assoc.* **90** 773–95
[25] Kelley C T 1995 *Iterative Methods for Linear, Nonlinear Equations* (Philadelphia: SIAM)
[26] Khan M E, Mohamed S and Murphy K 2012 Fast Bayesian inference for nonconjugate Gaussian process regression *Neural Information Processing Systems* (New York: American Institute of Physics) pp 3140–8
[27] Kulis B, Sustik M A and Dhillon I S 2009 Low-rank kernel learning with Bregman matrix divergences *J. Mach. Learn. Res.* **10** 341–76
[28] Kullback S and Leibler R A 1951 On information and sufficiency *Ann. Math. Stat.* **22** 79–86
[29] Lu Y, Stuart A M and Weber H 2017 Gaussian approximations for probability measures on $\boldsymbol{R}^d$ *SIAM/ASA J. Uncertain. Quantif.* **5** 1136–65
[30] Minka T P 2001 Expectation propagation for approximate Bayesian inference *Proc. 17th Conf. Uncertainty in Artificial Intelligence* (San Francisco, CA: Morgan Kaufmann Publishers Inc.)
[31] Opper M and Archambeau C 2009 The variational Gaussian approximation revisited *Neural Comput.* **21** 786–92
[32] Ormerod J T and Wand M P 2012 Gaussian variational approximate inference for generalized linear mixed models *J. Comput. Graph. Stat.* **21** 2–17
[33] Pillow J 2007 Likelihood-based approaches to modeling the neural code *Bayesian Brain: Probabilistic Approaches to Neural Coding* ed K Doya *et al* (Cambridge: MIT Press) pp 53–70
[34] Pinski F J, Simpson G, Stuart A M and Weber H 2015 Kullback–Leibler approximation for probability measures on infinite-dimensional spaces *SIAM J. Math. Anal.* **47** 4091–122
[35] Ravikumar P, Wainwright M J, Raskutti G and Yu B 2011 High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence *Electron. J. Stat.* **5** 935–80
[36] Rohde D and Wand M P 2016 Semiparametric mean field variational Bayes: general principles and numerical issues *J. Mach. Learn. Res.* **17** 47
[37] Schuster T, Kaltenbacher B, Hofmann B and Kazimierski K S 2012 *Regularization Methods in Banach Spaces* (Berlin: de Gruyter & Co)
[38] Stuart A M 2010 Inverse problems: a Bayesian perspective *Acta Numer.* **19** 451–559
[39] Tierney L and Kadane J B 1986 Accurate approximations for posterior moments and marginal densities *J. Am. Stat. Assoc.* **81** 82–6
[40] Wainwright M J and Jordan M I 2008 Graphical models, exponential families, and variational inference *Found. Trends© Mach. Learn.* **1** 1–305
[41] Yavuz M and Fessler J A 1997 New statistical models for randoms-precorrected PET scans *Information Processing in Medical Imaging* (*Lecture Notes in Computer Science* vol 1230) (Berlin: Springer)
[42] Zhang Y and Leithead W E 2007 Approximate implementation of the logarithm of the matrix determinant in Gaussian process regression *J. Stat. Comput. Simul.* **77** 329–48