Journal of Machine Learning Research 17 (2016) 1-47 Submitted 6/15; Revised 5/16; Published 9/16

Semiparametric Mean Field Variational Bayes: General Principles and Numerical Issues

David Rohde DAVIDJROHDE@GMAIL.COM
School of Mathematical and Physical Sciences
University of Techonology Sydney
P.O. Box 123, Broadway, 2007, Australia

Matt P. Wand MATT.WAND@UTS.EDU.AU
School of Mathematical and Physical Sciences
University of Techonology Sydney
P.O. Box 123, Broadway, 2007, Australia

Editor: Xiaotong Shen

Abstract

We introduce the term *semiparametric mean field variational Bayes* to describe the relaxation of mean field variational Bayes in which some density functions in the product density restriction are pre-specified to be members of convenient parametric families. This notion has appeared in various guises in the mean field variational Bayes literature during its history and we endeavor to unify this important topic. We lay down a general framework and explain how previous relevant methodologies fall within this framework. A major contribution is elucidation of numerical issues that impact semiparametric mean field variational Bayes in practice.

Keywords: Bayesian Computing, Factor Graph, Fixed-form Variational Bayes, Fixed-point Iteration, Non-conjugate Variational Message Passing, Nonlinear Conjugate Gradient Method

1. Introduction

We expound *semiparametric mean field variational Bayes*, a powerful combination of the notions of minimum Kullback-Leibler divergence and mean field restriction, that allows fast and often accurate approximate Bayesian inference for a wide range of scenarios. Most of its foundational literature and applications are in Machine Learning. However, semiparametric mean field variational Bayes is also an important paradigm for Statistics in the age of very big sample sizes and models.

Several articles concerned with deterministic approximate Bayesian inference, such as Barber and Bishop (1997), Honkela et al. (2010), Knowles and Minka (2011), Tan and Nott (2013), Wand (2014) and Menictas and Wand (2015), have demonstrated that modification of mean field variational Bayes (e.g. Wainwright and Jordan, 2008) to include pre-specified parametric families in the product density posterior approximation can have great practical benefits. For example, Barber and Bishop (1997) used a pre-specified Multivariate Normal distribution for the posterior density of the vector of adaptive parameters in multilayer neural networks while Tan and Nott (2013) derived a closed form variational approximate

algorithm for Bayesian Poisson mixed models by pre-specifying the fixed and random effects parameters to have Multivariate Normal distributions. Knowles and Minka (2011) took a message passing approach to mean field variational Bayes and explain how their approach to inclusion of pre-specified parametric families allows modular inference algorithms for arbitrary factors. Some recent articles on this topic have used the terms *fixed-form variational Bayes* (Honkela et al., 2010) and *non-conjugate variational message passing* (Knowles & Minka, 2011), to describe this modification of mean field variational Bayes. However, in this article we argue for adoption of the term *semiparametric mean field variational Bayes*.

Although we give a precise mathematical description of semiparametric mean field variational Bayes in Section 2, it simply refers to the relaxation of ordinary mean field variational Bayes in which some of the density functions in the postulated product density form are pre-specified to be particular parametric density functions, often chosen for reasons of tractability. This is a 'halfway house' between fully parametric approximation of the joint posterior density function of the model parameters with minimum Kullback-Leibler divergence used for parameter choice and pure mean field variational Bayes in which there is no parametric specification at all – only the product restriction. The following comments apply to our general framework:

- Semiparametric mean field variational Bayes is a modification of mean field variational Bayes that could be carried out via a message passing approach, as done by Knowles and Minka (2011), or by using the more common q-density approach used in, for example, Bishop (2006) and Ormerod and Wand (2010).

- The notion of conjugacy is not intrinsic to semiparametric mean field variational Bayes. The principle may be applied regardless of conjugacy relationships amongst the messages and/or q-densities. Therefore, the 'non-conjugacy' label used in recent articles for semiparametric relaxations of mean field variational Bayes is somewhat misleading.

- Contributions such as Knowles and Minka (2011) and Tan and Nott (2013) restrict attention to pre-specification of parametric families that are of exponential family form (e.g. Wainwright and Jordan, 2008). Whilst exponential family density functions have tractability advantages when used in semiparametric mean field variational Bayes, there is no intrinsic reason for only such densities to be used. In Section 5 we illustrate this point using pre-specified Skew-Normal density functions, which are not within the exponential family.

- Recent articles on non-conjugate variational message passing, such as Knowles and Minka (2011), Tan and Nott (2013) and Menictas and Wand (2015) used fixed-point iteration to minimize the Kullback-Leibler divergence or, equivalently, maximize the lower bound on the marginal log-likelihood. Theorem 1 of Knowles and Minka (2011) constitutes such an approach. However, any optimization approach could be used for obtaining the Kullback-Leibler optimal parameters such as Newton-Raphson iteration, quasi-Newton iteration, stochastic gradient descent, the Nelder-Mead simplex method and various hybrids and modifications of such methods.

- Some articles, such as the recent Challis and Barber (2013), are concerned solely with approximate inference via minimum divergence according to the pre-specification that

the posterior is within a parametric family such as Multivariate Normal with banded Cholesky covariance matrix factors. These contributions represent special cases of semiparametric mean field variational Bayes and their findings have relevance to the more general situation.

The main purposes of this article are:

(1) Bring together the literature on semiparametric mean field variational Bayes and identify its core tenets.

(2) Lay out and discuss numerical issues that arise in semiparametric mean field variational Bayes, which have a significant practical implications for this body of methodology.

The resulting exposition constitutes the first compendium on semiparametric mean field variational Bayes at its fullest level of generality. It can also be used as a basis for enhancements of semiparametric mean field variational Bayes methodology.

We use two examples to elucidate the general principles and numerical issues. The first, Example 1, involves a Bayesian model with a single parameter and, hence, is such that mean field approximation is not required. The simplicity of Example 1 allows a deep appreciation of the various issues with minimal notational overhead. Example 2 is the Bayesian Poisson mixed model treated in Wand (2014) and benefits from semiparametric mean field variational Bayes methodology. It demonstrates issues with high-dimensional optimization problems that are intrinsic to practical implementation.

One of the main outcomes of our numerical investigations is that fitting exponential family density functions via *natural* fixed-point iteration has some attractive properties. By 'natural', we mean a simple version of fixed-point iteration that arises when natural parametrizations are used. As we explain in Section 4, natural fixed-point iterations use *Riemannian* gradients to step through the parameter space, which is generally more efficient than ordinary gradients. The benefits of Riemannian gradient-based algorithms for Machine Learning problems goes back at least to Amari (1998). Such algorithms are the basis of the semiparametric mean field variational Bayes approach of Honkela et al. (2010).

In Section 2 we describe semiparametric mean field variational Bayes in full generality. A general overview of optimization strategies, pertinent to semiparametric mean field variational Bayes, is given in Section 3. The important special case of pre-specified exponential family density functions is treated in Section 4. Section 5 deals with the more difficult non-exponential family case via an illustrative example. Some closing remarks are given in Section 6.

2. General Principles

Semiparametric mean field variational Bayes is an approximate Bayesian inference method based on the principle of minimum Kullback-Leibler divergence. For arbitrary density functions p_1 and p_2 on $\mathbb{R}^d$,

$$\mathrm{KL}(p_1 \,\|\, p_2) \equiv \int_{\mathbb{R}^d} p_1(\boldsymbol{x}) \, \log\left\{p_1(\boldsymbol{x})/p_2(\boldsymbol{x})\right\} d\boldsymbol{x}$$

denotes the *Kullback-Leibler divergence* of p_2 from p_1. Note that

$$\mathrm{KL}(p_1 \,\|\, p_2) \geq 0 \quad \text{for any } p_1 \text{ and } p_2. \tag{1}$$

Consider a generic Bayesian model with observed data $\boldsymbol{\mathscr{D}}$ and parameter vector $(\boldsymbol{\theta}, \boldsymbol{\phi})$. The reason for this notational decomposition of the parameter vector will soon become apparent. Throughout this section we assume that $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and $\boldsymbol{\mathscr{D}}$ are continuous random vectors with density functions $p(\boldsymbol{\theta}, \boldsymbol{\phi})$ and $p(\boldsymbol{\mathscr{D}})$. The situation where some components are discrete has similar treatment with summations replacing integrals. Bayesian inference for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is based on the posterior density function

$$p(\boldsymbol{\theta}, \boldsymbol{\phi} | \boldsymbol{\mathscr{D}}) = \frac{p(\boldsymbol{\mathscr{D}}, \boldsymbol{\theta}, \boldsymbol{\phi})}{p(\boldsymbol{\mathscr{D}})}.$$

The denominator, $p(\boldsymbol{\mathscr{D}})$, is usually referred to as the *marginal likelihood* or the *model evidence*.

Let $q(\boldsymbol{\theta}, \boldsymbol{\phi})$ be an arbitrary density function over the parameter space of $(\boldsymbol{\theta}, \boldsymbol{\phi})$. The essence of variational approximate inference is to restrict $q(\boldsymbol{\theta}, \boldsymbol{\phi})$ to some class of density functions $\mathcal{Q}$ and then use the optimal q-density function, given by

$$q^*(\boldsymbol{\theta}, \boldsymbol{\phi}) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \, \mathrm{KL}\Big\{ q(\boldsymbol{\theta}, \boldsymbol{\phi}) \,\Big\|\, p(\boldsymbol{\theta}, \boldsymbol{\phi} | \boldsymbol{\mathscr{D}}) \Big\}, \tag{2}$$

as an approximation to the posterior density function $p(\boldsymbol{\theta}, \boldsymbol{\phi} | \boldsymbol{\mathscr{D}})$.

Simple algebraic arguments (e.g. Section 2.1 of Ormerod and Wand, 2010) lead to

$$\log \, p(\boldsymbol{\mathscr{D}}) = \mathrm{KL}\Big\{ q(\boldsymbol{\theta}, \boldsymbol{\phi}) \,\Big\|\, p(\boldsymbol{\theta}, \boldsymbol{\phi} | \boldsymbol{\mathscr{D}}) \Big\} + \log \underline{p}(\boldsymbol{\mathscr{D}}; q) \tag{3}$$

where

$$\underline{p}(\boldsymbol{\mathscr{D}}; q) \equiv \exp \left[\int \int q(\boldsymbol{\theta}, \boldsymbol{\phi}) \, \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})}{q(\boldsymbol{\theta}, \boldsymbol{\phi})} \right\} \, d\boldsymbol{\theta} \, d\boldsymbol{\phi} \right]. \tag{4}$$

From (1) we have

$$\underline{p}(\boldsymbol{\mathscr{D}}; q) \leq p(\boldsymbol{\mathscr{D}}) \text{ for any } q(\boldsymbol{\theta}, \boldsymbol{\phi})$$

showing that $\underline{p}(\boldsymbol{\mathscr{D}}; q)$ is a lower bound on the marginal likelihood. The non-negativity condition (1) means that an equivalent form for the optimal q-density function is

$$q^*(\boldsymbol{\theta}, \boldsymbol{\phi}) = \underset{q \in \mathcal{Q}}{\operatorname{argmax}} \, \underline{p}(\boldsymbol{\mathscr{D}}; q). \tag{5}$$

This alternative optimization problem has the attractive interpretation of $q^*(\boldsymbol{\theta}, \boldsymbol{\phi})$ being chosen to maximize a lower bound on the marginal likelihood. For the remainder of this article we work with (5) rather than (2).

Parametric variational approximate inference involves setting

$$\mathcal{Q} = \{ q(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi \},$$

corresponding to a parametric family of density functions with parameter vector $\boldsymbol{\xi}$ ranging over Ξ. In this case (5) reduces to

$$q^*(\boldsymbol{\theta}, \boldsymbol{\phi}) = \underset{\xi \in \Xi}{\operatorname{argmax}} \, \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi}), \tag{6}$$

where $\underline{p}(\mathscr{D}; q, \boldsymbol{\xi})$ is the marginal likelihood lower bound defined by (4), but with the dependence on $\boldsymbol{\xi}$ reflected in the notation.

An early contribution of this type is Hinton and van Camp (1993) who used minimum Kullback-Leibler divergence for Gaussian approximation of posterior density functions in neural networks models. Gaussian $\mathcal{Q}$ families have also been used by Lappalainen and Honkela (2000), Archambeau et al. (2007), Raiko et al. (2007), Nickisch and Rasmussen (2008), Honkela and Valpola (2005), Honkela et al. (2007), Honkela et al. (2008) and Opper and Archambeau (2009). The recent contribution by Challis and Barber (2013) is an in-depth coverage of Gaussian minimum Kullback-Leibler approximate inference. Salimans and Knowles (2013) devised a stochastic approximation algorithm for solving (6) when $\mathcal{Q}$ is a parametric family of exponential family form. Gershman et al. (2012) and Zobay (2014) investigated Gaussian-mixture extensions.

In what one may label a *nonparametric* variational approximation approach, ordinary mean field variational Bayes uses restricted q-density spaces such as

$$\mathcal{Q} = \{q(\boldsymbol{\theta}, \boldsymbol{\phi}) : q(\boldsymbol{\theta}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}_1) \cdots q(\boldsymbol{\theta}_M)\, q(\boldsymbol{\phi})\} \text{ for some partition } \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \text{ of } \boldsymbol{\theta}. \quad (7)$$

The word 'nonparametric' is justified by the fact that there is no pre-specification that the q-density, or any of its factors, belong to a particular parametric family. Restriction of $q(\boldsymbol{\theta}, \boldsymbol{\phi})$ to a product density form is the only pre-specification being made. An iterative scheme for solving (5) under (7) follows from the last displayed equation given in Section 10.1.1 of Bishop (2006). The scheme is listed explictly as Algorithm 1 of Ormerod and Wand (2010). Note that a simple adjustment that caters for $(\boldsymbol{\theta}, \boldsymbol{\phi})$, rather than $\boldsymbol{\theta}$, is required for notation being used here. Gershman et al. (2012) also use the word 'nonparametric' to describe a variational approximation approach. However, their methodology is parametric in the sense of the terminology that we are using here.

We propose that the term *semiparametric mean field variational Bayes* be used for restrictions of the form:

$$\mathcal{Q} = \{q(\boldsymbol{\theta}, \boldsymbol{\phi}) : q(\boldsymbol{\theta}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}_1) \cdots q(\boldsymbol{\theta}_M)\, q(\boldsymbol{\phi}; \boldsymbol{\xi}),\ \boldsymbol{\xi} \in \Xi\} \quad (8)$$

where $\{q(\boldsymbol{\phi}; \boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi\}$ is a pre-specified parametric family of density functions in $\boldsymbol{\phi}$. Under (8) there is no insistence on the $q(\boldsymbol{\theta}_i)$ having a particular parametric form. For models possessing particular conjugacy properties the optimal q-densities, $q^*(\boldsymbol{\theta}_i)$, will belong to relevant conjugate families. However, in general, the optimal q-densities of the θ_i can assume arbitrary forms; see, for example, Figure 6 of Pham et al. (2013). The quality of a variational approximation is limited by the restrictions imposed by the particular choice of $\mathcal{Q}$. Semiparametric mean field variational Bayes imposes a product density restriction and then a parametric constraint on one of the factors. The overall quality of the approximation is determined by the combination of these two restrictions. While the estimated nonparametric factors are optimal given the product restriction, a parametric restriction with fewer product assumptions may be more accurate.

We now turn to the practical problem of solving the optimization problem (5) when the q-density restriction is of the form (8). Appropriate strategies for solving (5) depend on the nature of $q(\phi; \boldsymbol{\xi})$ as a function of $\boldsymbol{\xi}$ and the set Ξ. Some possibilities are:

(A) Ξ is a finite set.

(B) Ξ is an open subset of $\mathbb{R}^d$ for some $d \in \mathbb{N}$ and $q(\phi; \boldsymbol{\xi})$ is smooth function of $\boldsymbol{\xi}$ over $\boldsymbol{\xi} \in \Xi$.

(C) Ξ is an open subset of $\mathbb{R}^d$ for some $d \in \mathbb{N}$ and $q(\phi; \boldsymbol{\xi})$ is a non-smooth function of $\boldsymbol{\xi}$ over $\boldsymbol{\xi} \in \Xi$.

(D) Ξ is a complicated set that does not satisfy any of the descriptions given in (A)–(C).

For the vast majority of models in common use and $q(\phi; \boldsymbol{\xi})$ families (B) applies and most of the remainder of this article is devoted to that case. However, we will first briefly deal with (A) in Section 2.1, since it aids understanding of the semiparametric extension of mean field variational Bayes. To date, we are unaware of any semiparametric mean field variational Bayes contributions where (C) or (D) apply, so these cases are left aside.

2.1 Finite Parameter Space Case

Suppose that Ξ is a finite set. Then Algorithm 1 is the natural extension of the mean field variational Bayes coordinate ascent algorithm given, for example, in Section 10.1.1 of Bishop (2006) and Algorithm 1 of Ormerod and Wand (2010). In Algorithm 1, and elsewhere, the notation $\boldsymbol{\theta} \backslash \boldsymbol{\theta}_i$ denotes the vector $\boldsymbol{\theta}$ with the entries of $\boldsymbol{\theta}_i$ excluded.

For each $\boldsymbol{\xi} \in \Xi$:

Initialize: $q(\boldsymbol{\theta}_2), \ldots, q(\boldsymbol{\theta}_M)$.

Cycle:

$$q(\boldsymbol{\theta}_1) \leftarrow \frac{\exp\left[E_{q(\boldsymbol{\theta}\backslash\boldsymbol{\theta}_1)\,q(\boldsymbol{\phi};\boldsymbol{\xi})}\{\log\, p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})\}\right]}{\int \exp\left[E_{q(\boldsymbol{\theta}\backslash\boldsymbol{\theta}_1)\,q(\boldsymbol{\phi};\boldsymbol{\xi})}\{\log\, p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})\}\right] d\boldsymbol{\theta}_1}$$

$$\vdots$$

$$q(\boldsymbol{\theta}_M) \leftarrow \frac{\exp\left[E_{q(\boldsymbol{\theta}\backslash\boldsymbol{\theta}_M)\,q(\boldsymbol{\phi};\boldsymbol{\xi})}\{\log\, p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})\}\right]}{\int \exp\left[E_{q(\boldsymbol{\theta}\backslash\boldsymbol{\theta}_M)\,q(\boldsymbol{\phi};\boldsymbol{\xi})}\{\log\, p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})\}\right] d\boldsymbol{\theta}_M}$$

until the increase in $\underline{p}(\mathscr{D}; q, \boldsymbol{\xi})$ is negligible.

$q^*(\boldsymbol{\theta}_i; \boldsymbol{\xi}) \leftarrow q(\boldsymbol{\theta}_i),\ 1 \le i \le M \quad ; \quad \underline{p}(\mathscr{D}; q^*, \boldsymbol{\xi}) \leftarrow \underline{p}(\mathscr{D}; q, \boldsymbol{\xi})$.

$$\boldsymbol{\xi}^* \leftarrow \operatorname*{argmax}_{\boldsymbol{\xi}\in\Xi} \underline{p}(\mathscr{D}; q^*, \boldsymbol{\xi}) \quad ; \quad q^*(\boldsymbol{\theta}_i) \leftarrow q^*(\boldsymbol{\theta}_i; \boldsymbol{\xi}^*),\ 1 \le i \le M.$$

Algorithm 1: *Coordinate ascent algorithm for semiparametric mean field variational Bayes when Ξ is a finite parameter space.*

For each value of $\boldsymbol{\xi}$ in Ξ, Algorithm 1 is essentially the ordinary mean field variational Bayes iterative algorithm — but with the density function of ϕ set to the parametric density function $q(\phi; \boldsymbol{\xi})$. The optimal $\boldsymbol{\xi}$ is then found by maximizing over the approximate marginal likelihood values that are recorded for each element of Ξ.

2.2 Infinite Parameter Space Case

Algorithm 1 shows how to obtain the Kullback-Leibler-optimal $q(\theta_i)$ and $q(\phi; \boldsymbol{\xi})$ densities in the case where Ξ is finite. However, for common parametric families such as the Normal and Gamma, Ξ is infinite and the solution of (5) under (8) is more delicate. The coordinate ascent idea used to obtain the $q^*(\theta_i)$ in Algorithm 1 can still be entertained. However, it needs to be combined with an optimization scheme that searches for the optimal $\boldsymbol{\xi}$ over the infinite space Ξ.

For the remainder of this article we focus on the problem of solving (5) under restriction (B) on the q-density parameter space Ξ. We start by studying the criterion function $\underline{p}(\mathscr{D}; q, \boldsymbol{\xi})$ and special forms that it takes under the mean field restriction. The notions of entropy and factor graphs are shown to be very relevant and useful. We then introduce two running examples, Example 1 and Example 2, to illustrate the issues involved. Since Example 1 has only one parameter requiring inference, this is not a fully-fledged semiparametric mean field variational Bayes problem and the optimization problem is of the form (6). Additionally, (6) for Example 1 is a bivariate optimization problem which allows deeper probing of the numerical analytic issues. Example 2 uses the Poisson mixed model, treated in Section 5.1 of Wand (2014), as our main semiparametric mean field variational Bayes example. It is substantial enough to convey various practical issues but also has a closed form $\log \underline{p}(q; \boldsymbol{\xi})$ expression that allows purely algebraic exposition.

2.2.1 Entropy, Factor Graphs and the Marginal Log-Likelihood Lower Bound

If $\boldsymbol{x}$ is a random vector having density function p then the corresponding *entropy* is given by

$$\text{Entropy}(p) \equiv E_p\{-\log\, p(\boldsymbol{x})\}.$$

For many common distributional families, the entropy admits an algebraic expression in terms of the distribution's parameters. For example, if

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\}$$

is the Multivariate Normal density function of dimension d with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ then

$$\text{Entropy}(p; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \tfrac{1}{2}\, d\{1 + \log(2\pi)\} + \tfrac{1}{2} \log |\boldsymbol{\Sigma}|. \tag{9}$$

Note that the entropy is independent of the mean vector $\boldsymbol{\mu}$.

Another entropy expression, which arises in many Bayesian models and the example of Section 2.2.3, is that for the Inverse Gamma family of density functions. Let

$$x \sim \text{Inverse-Gamma}(\kappa, \lambda)$$

denote the random variable x having density function

$$p(x; \kappa, \lambda) = \frac{\kappa^\lambda}{\Gamma(\kappa)}\, x^{-\kappa - 1} \exp(-\lambda / x), \quad x > 0,$$

with parameters $\kappa, \lambda > 0$. In this case

$$\text{Entropy}(p; \kappa, \lambda) = \log(\lambda) + \kappa + \log\{\Gamma(\kappa)\} - (\kappa + 1)\text{digamma}(\kappa) \tag{10}$$

where $\text{digamma}(x) \equiv (d/dx) \log \Gamma(x)$ is the digamma function.

The next relevant concept is that of a *factor graph*, which we first explain via an example. Consider the approximate Bayesian inference problem according to the semiparametric mean field variational Bayes restriction (8). Figure 1 is the factor graph for an $M = 9$ example of (8) with the joint density function of all random vectors in the model factorizing as follows:

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) \quad = \quad & f_1(\boldsymbol{\theta}_1)\, f_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\phi})\, f_3(\boldsymbol{\theta}_3)\, f_4(\boldsymbol{\theta}_4)\, f_5(\boldsymbol{\theta}_5, \boldsymbol{\phi})\, f_6(\boldsymbol{\theta}_5)\, f_7(\boldsymbol{\theta}_6, \boldsymbol{\phi})\, f_8(\boldsymbol{\theta}_6) \\ & \times f_9(\boldsymbol{\theta}_7, \boldsymbol{\theta}_8, \boldsymbol{\theta}_9)\, f_{10}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4) \end{aligned} \tag{11}$$

for *factors* $f_1, \ldots, f_{10}$. Note that some of these factors depend on the data vector $\boldsymbol{x}$, but the dependence is suppressed in the f_j notation. Specific examples are given in Sections 2.2.2 and 2.2.3. In Figure 1 the circles correspond to the components of the mean field product restriction

$$q(\boldsymbol{\phi}) \prod_{i=1}^{9} q(\boldsymbol{\theta}_i) \tag{12}$$

and are called *stochastic nodes*. The solid squares correspond to the factors f_j, $1 \leq j \leq 10$, and are called *factor nodes*. Edges join each factor node f_j to those stochastic nodes that are included in the f_j function.

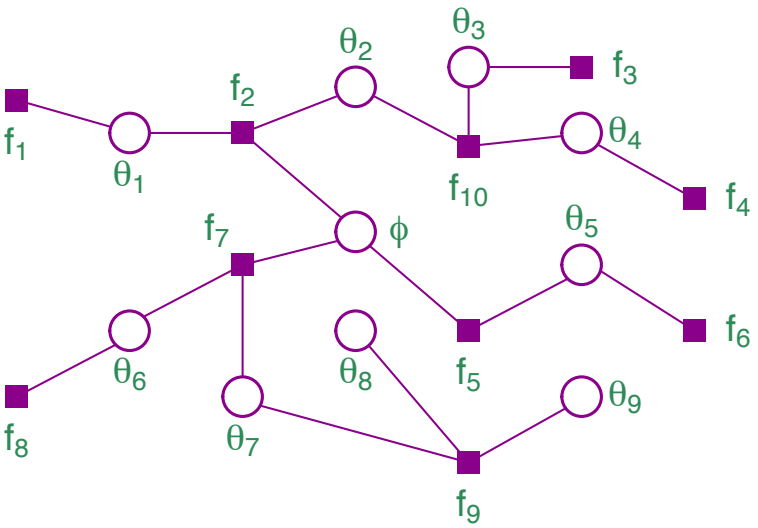

Figure 1: *Factor graph corresponding to the model (11) and q-density product restriction (12).*

Now consider the general case with semiparametric mean field restriction (8) and suppose that $p(\boldsymbol{x}, \boldsymbol{\theta}, \phi)$ has N factors f_j, $1 \leq j \leq N$. Then the marginal log-likelihood lower bound admits the following expression in terms of the components of the corresponding factor graph:

$$\log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi}) = \text{Entropy}\{q(\phi; \boldsymbol{\xi})\} + \sum_{i=1}^{M} \text{Entropy}\{q(\boldsymbol{\theta}_i)\} + \sum_{j=1}^{N} E_q\{\log(f_j)\}.$$

The $\boldsymbol{\phi}$-*localized* component of $\log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})$, which we denote by $\log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]}$, is

$$\log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]} \equiv \text{Entropy}\{q(\boldsymbol{\phi}; \boldsymbol{\xi})\} + \sum_{j \in \text{ neighbors}(\boldsymbol{\phi})} E_q\{\log(f_j)\} \tag{13}$$

where

$$\begin{aligned} \text{neighbors}(\boldsymbol{\phi}) &\equiv \{1 \leq j \leq N : f_j \text{ is a neighbor of } \boldsymbol{\phi} \text{ on the factor graph}\} \\ &= \{1 \leq j \leq N : f_j \text{ involves } \boldsymbol{\phi}\}. \end{aligned}$$

For the factor graph shown in Figure 1 $\text{neighbors}(\boldsymbol{\phi}) = \{2, 5, 7\}$ and so have

$$\log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]} = \text{Entropy}\{q(\boldsymbol{\phi}; \boldsymbol{\xi})\} + E_q\{\log(f_2)\} + E_q\{\log(f_5)\} + E_q\{\log(f_7)\}.$$

as the $\boldsymbol{\phi}$-localized component of $\log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})$.

For large Bayesian models it is prudent to maximize this localized approximate log-likelihood as part of a coordinate ascent scheme involving all q-density parameters. Such an approach, combined with the *locality property* of mean field variational Bayes (e.g. Wand et al., 2011, Section 3), allows for streamlined handling of arbitrarily large models. We formalize this approach to semiparametric mean field variational Bayes in the shape of Algorithm 2 in the upcoming Section 2.2.4. However, we first give some concrete examples of mean field variational Bayes with pre-specified parametric family q-density functions.

2.2.2 Example 1: Gumbel Random Sample

A Bayesian model for a random sample $x_1, \ldots, x_n$ from a Gumbel distribution with location parameter ϕ and unit scale is

$$p(x_1, \ldots, x_n \,|\, \phi) = \prod_{i=1}^{n} \exp\{-(x_i - \phi) - e^{-(x_i - \phi)}\}, \quad \phi \sim N(\mu_\phi, \sigma_\phi^2). \tag{14}$$

The posterior density function of ϕ is

$$p(\phi|\boldsymbol{x}) = \frac{\exp\left\{n\,\phi - e^{\phi} \sum_{i=1}^{n} e^{-x_i} - \frac{1}{2\sigma_\phi^2}(\phi - \mu_\phi)^2\right\}}{\int_{-\infty}^{\infty} \exp\left\{n\,\phi' - e^{\phi'} \sum_{i=1}^{n} e^{-x_i} - \frac{1}{2\sigma_\phi^2}(\phi' - \mu_\phi)^2\right\} d\phi'}. \tag{15}$$

The denominator on the right-hand side of (15) is not available in closed form. This implies that numerical methods such as quadrature are required to obtain the Bayes estimate of ϕ

and corresponding credible sets. Instead we consider minimum Kullback-Leibler divergence approximation of $p(\phi|\boldsymbol{x})$ over a parametric pre-specified family. Let

$$\mathcal{Q} = \{q(\phi;\boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi\}$$

be such a parametric family. Then the optimal q-density is $q(\phi;\boldsymbol{\xi}^*)$ where

$$\boldsymbol{\xi}^* = \underset{\boldsymbol{\xi}\in\Xi}{\operatorname{argmax}}\, \underline{p}(q;\boldsymbol{\xi}). \tag{16}$$

Figure 2 shows the factor graph of the model, with factors $p(\phi)$ and $p(\boldsymbol{x}|\phi)$ neighboring the stochastic node ϕ.

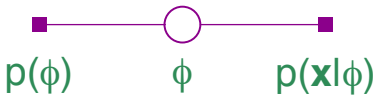

Figure 2: *Factor graph for the Example 1 model.*

The marginal log-likelihood lower bound is

$$\log \underline{p}(\boldsymbol{x};\boldsymbol{\xi}) = \text{Entropy}\{q(\phi;\boldsymbol{\xi})\} + E_q\{\log\, p(\phi)\} + E_q\{\log\, p(\boldsymbol{x}|\phi)\} \tag{17}$$

and the contributions to $\log \underline{p}(\boldsymbol{x};\boldsymbol{\xi})$ from the factors are

$$\begin{aligned} E_q\{\log\, p(\phi)\} &= -\tfrac{1}{2\sigma_\phi^2}\Big[\{E_{q(\phi;\boldsymbol{\xi})}(\phi) - \mu_\phi\}^2 + \text{Var}_{q(\phi;\boldsymbol{\xi})}(\phi)\Big] \\ \text{and}\quad E_q\{\log\, p(\boldsymbol{x}|\phi)\} &= n\, E_{q(\phi;\boldsymbol{\xi})}(\phi) - M_{q(\phi;\boldsymbol{\xi})}(1)\, \textstyle\sum_{i=1}^n e^{-x_i} - n\overline{x} \end{aligned} \tag{18}$$

where $M_{q(\phi;\boldsymbol{\xi})}$ is the moment generating function corresponding to $q(\phi;\, \boldsymbol{\xi})$.

Now suppose that

$$\mathcal{Q} = \left\{ q(\phi;\mu_{q(\phi)},\sigma^2_{q(\phi)}) = \frac{1}{\sqrt{2\pi\sigma^2_{q(\phi)}}}\, \exp\left\{\frac{-(\phi-\mu_{q(\phi)})^2}{2\sigma^2_{q(\phi)}}\right\} : \mu_{q(\phi)} \in \mathbb{R}, \sigma^2_{q(\phi)} > 0 \right\}$$

corresponding to the Normal family with q-density parameter vector $\boldsymbol{\xi} = (\mu_{q(\phi)}, \sigma^2_{q(\phi)})$ and parameter space $\Xi = \mathbb{R}\times\mathbb{R}_+$ where $\mathbb{R}_+ \equiv (0,\infty)$ is the positive half-line. Then, from the entropy result (9) and the well-known expression for the moment generating function of the Normal distribution we obtain

$$\begin{aligned} \log \underline{p}(\boldsymbol{x};\mu_{q(\phi)},\sigma^2_{q(\phi)}) = {} & \tfrac{1}{2}\{1+\log(2\pi)\} + \tfrac{1}{2}\log(\sigma^2_{q(\phi)}) + n\,\mu_{q(\phi)} \\ & - \exp(\mu_{q(\phi)} + \tfrac{1}{2}\sigma^2_{q(\phi)}) \sum_{i=1}^n e^{-x_i} - \tfrac{1}{2\sigma_\phi^2}\{(\mu_{q(\phi)}-\mu_\phi)^2 + \sigma^2_{q(\phi)}\} - n\,\overline{x}. \end{aligned}$$

It follows that the Kullback-Leibler optimal Normal q-density function is $q(\phi;\mu^*_{q(\phi)},(\sigma^2_{q(\phi)})^*)$ where

$$\begin{bmatrix} \mu^*_{q(\phi)} \\ (\sigma^2_{q(\phi)})^* \end{bmatrix} = \underset{\mu_{q(\phi)}\in\mathbb{R},\sigma^2_{q(\phi)}>0}{\operatorname{argmax}} \left\{ f^N_{\text{Ex1}}\left(\mu_{q(\phi)},\sigma^2_{q(\phi)}; n, \sum_{i=1}^n e^{-x_i}, \mu_\phi, \sigma^2_\phi\right)\right\} \tag{19}$$

and

$$f^N_{\text{Ex1}}(x, y; a, b, c, d) = \tfrac{1}{2}\log(y) + a\,x - b\,\exp(x + \tfrac{1}{2}\,y) - \tfrac{1}{2d}\{(x-c)^2 + y\}. \tag{20}$$

The main arguments satisfy $x \in \mathbb{R}$, $y > 0$ and the auxiliary arguments are such that $a, b, d > 0$ and $c \in \mathbb{R}$. From (19) we see that the minimum Kullback-Leibler divergence problem (16), where $\mathcal{Q}$ is the Normal family, reduces to a non-linear bivariate optimization problem. Theory given in Challis and Barber (2013) applies to this example. For example, results given in their Section 3.2 can be used to establish that $f^N_{\text{Ex1}}(x, y; a, b, c, d)$ is jointly concave in x and $\sqrt{y}$.

In Section 3 we study strategies for solving such problems and apply them to this example in Section 4.2.

2.2.3 Example 2: Poisson Mixed Model

A single variance component *Poisson mixed model* is

$$\begin{gathered} y_i \,|\, \boldsymbol{\beta}, \boldsymbol{u} \overset{\text{ind.}}{\sim} \text{Poisson}[\exp\{(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_i\}], \quad 1 \le i \le n, \\ \boldsymbol{u} \,|\, \sigma^2 \sim N(0, \sigma^2 \boldsymbol{I}_K), \quad \sigma^2 \,|\, a \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a), \\ \boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma^2_{\boldsymbol{\beta}}\, \boldsymbol{I}_p), \quad a \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A^2) \end{gathered} \tag{21}$$

where $\boldsymbol{X}$ is an $n \times p$ fixed effects design matrix, $\boldsymbol{Z}$ is an $n \times K$ random effects design matrix. Note that the prior on σ in (21) is the Half Cauchy distribution with scale parameter A:

$$p(\sigma) = \frac{(2/\pi)}{A\{1 + (\sigma/A)^2\}}, \quad \sigma > 0.$$

In (21) $\sigma_{\boldsymbol{\beta}} > 0$ and $A > 0$ are hyperparameters to be chosen by the analyst.

A mean field approximation to the joint posterior density function of the model parameters is

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, a \,|\, \boldsymbol{y}) \;\approx\; q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\sigma^2)\, q(a). \tag{22}$$

As detailed in Appendix A.3, the optimal q-density functions satisfy

$$\begin{gathered} q^*(\sigma^2) \text{ and } q^*(a) \text{ are both Inverse-Gamma density functions, and} \\ q^*(\boldsymbol{\beta}, \boldsymbol{u}) \propto \exp\big\{\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) - \boldsymbol{1}^T \exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}) \\ -\tfrac{1}{2}\sigma^{-2}_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|^2 - \tfrac{1}{2}\, E_{q(\sigma^2)}(1/\sigma^2)\|\boldsymbol{u}\|^2\big\}. \end{gathered} \tag{23}$$

Since $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ is not a standard form, numerical methods are required to obtain the variational approximate Bayes estimates and credible sets. *Semiparametric* mean field variational Bayes alternatives take the form

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, a \,|\, \boldsymbol{y}) \;\approx\; q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})\, q(\sigma^2)\, q(a) \tag{24}$$

where $\{q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi\}$ is a pre-specified parametric family of density functions. The optimal density functions $q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi}^*)$, $q^*(\sigma^2)$ and $q^*(a)$ are found by minimizing

$$\text{KL}\Big\{q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})\, q(\sigma^2)\, q(a) \;\Big\|\; p(\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, a \,|\boldsymbol{y})\Big\}. \tag{25}$$

We now focus on solving (25).

In Appendix C of Wand (2014) it is shown that

$$q^*(\sigma^2) \text{ is an Inverse-Gamma}(\tfrac{1}{2}(K+1), B_{q(\sigma^2)}) \text{ density function, and}$$
$$q^*(a) \text{ is an Inverse-Gamma}(1, B_{q(a)}) \text{ density function}$$

where

$$B_{q(\sigma^2)} = \tfrac{1}{2}[\| E_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{u})\|^2 + \text{tr}\{\text{Cov}_{q(\beta,\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{u})\}] + \mu_{q(1/a)} \tag{26}$$

and $B_{q(a)} = \mu_{q(1/\sigma^2)} + A^{-2}$ with the definition

$$\mu_{q(1/v)} \equiv E_{q(v)}(1/v)$$

for a generic random variable v.

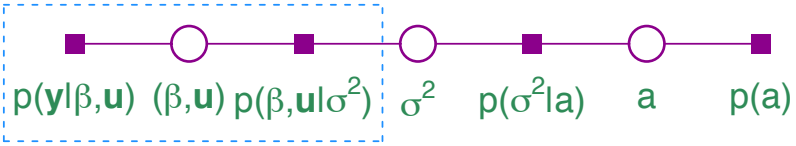

Figure 3: *Factor graph for the Example 2 model with stochastic nodes corresponding the mean field restriction (24). The dashed line box contains the stochastic node* $(\boldsymbol{\beta}, \boldsymbol{u})$ *and its neighboring factors.*

It remains to obtain the optimal value of $\boldsymbol{\xi}$ in $q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})$. Figure 3 shows the factor graph of the current model under mean field restriction (24). The lower bound on the marginal log-likelihood, in terms of the stochastic nodes and factors of Figure 3, is

$$\begin{aligned}\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\xi}) = &\text{Entropy}\{q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})\} + \text{Entropy}\{q(\sigma^2)\} + \text{Entropy}\{q(a)\} \\ &+E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u})\} + E_q\{\log p(\boldsymbol{\beta}, \boldsymbol{u}|\, \sigma^2)\} \\ &+E_q\{\log p(\sigma^2|\, a)\} + E_q\{\log p(a)\}.\end{aligned} \tag{27}$$

One could substitute (26) into the $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\xi})$ expression. This resulting marginal log-likelihood lower bound would depend on the q-densities only through $\boldsymbol{\xi}$ and could then be maximized over $\boldsymbol{\xi} \in \Xi$.

An alternative strategy, that scales better to larger models, is to use a coordinate ascent scheme that maximizes $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\xi})^{[(\boldsymbol{\beta},\boldsymbol{u})]}$, the $(\boldsymbol{\beta}, \boldsymbol{u})$-localized component of $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\xi})$, over $\boldsymbol{\xi} \in \Xi$. The relevant factors are those neighboring $(\boldsymbol{\beta}, \boldsymbol{u})$ in Figure 3, corresponding to the dashed line box. The quantity requiring maximization is then

$$\begin{aligned}\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\xi})^{[(\boldsymbol{\beta},\boldsymbol{u})]} &= \text{Entropy}\{q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})\} + E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u})\} + E_q\{\log p(\boldsymbol{\beta}, \boldsymbol{u}|\sigma^2)\} \\ &= \text{Entropy}\{q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})\} + \boldsymbol{y}^T\{\boldsymbol{X}\, E_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{\beta}) + \boldsymbol{Z}\, E_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{u})\} \\ &\quad -\mathbf{1}^T E_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\{\exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\} \\ &\quad -\tfrac{1}{2\sigma_{\boldsymbol{\beta}}^2}\left[\| E_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{\beta})\|^2 + \text{tr}\{\text{Cov}_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{\beta})\}\right] \\ &\quad -\tfrac{1}{2}\,\mu_{q(1/\sigma^2)}\left[\| E_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{u})\|^2 + \text{tr}\{\text{Cov}_{q(\beta,\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{u})\}\right] + \text{const}\end{aligned} \tag{28}$$

where 'const' denotes terms not depending on $\boldsymbol{\xi}$.

Next, suppose that $\mathcal{Q}$ corresponds to the family of Multivariate Normal density functions in $(\boldsymbol{\beta}, \boldsymbol{u})$:

$$q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) = (2\pi)^{-(p+K)/2} |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}|^{-1/2} \\ \times \exp\left\{ -\tfrac{1}{2} \left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \right)^T \boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\beta},\boldsymbol{u})} \left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} - \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \right) \right\}.$$

Then the $(\boldsymbol{\beta}, \boldsymbol{u})$-localized approximate marginal log-likelihood reduces to

$$\begin{aligned} \log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})^{[(\boldsymbol{\beta},\boldsymbol{u})]} = \; & \tfrac{1}{2} \log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| + \boldsymbol{y}^T \boldsymbol{C} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \\ & - \mathbf{1}^T \exp\left\{ \boldsymbol{C} \boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \tfrac{1}{2} \text{diagonal}(\boldsymbol{C} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{C}^T) \right\} \\ & - \tfrac{1}{2\sigma^2_{\boldsymbol{\beta}}} \left\{ \| \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}) \right\} \\ & - \tfrac{1}{2} \mu_{q(1/\sigma^2)} \left\{ \| \boldsymbol{\mu}_{q(\boldsymbol{u})} \|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u})}) \right\} + \text{const} \end{aligned} \tag{29}$$

where $\boldsymbol{C} \equiv [\boldsymbol{X} \; \boldsymbol{Z}]$, diagonal$(\boldsymbol{M})$ is the vector containing the diagonal entries of the square matrix $\boldsymbol{M}$, and $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ is the sub-vector of $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ corresponding to $\boldsymbol{\beta}$. Analogous definitions apply to $\boldsymbol{\mu}_{q(\boldsymbol{u})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{u})}$. Appendix A.3 provides derivational details for (29).

For this example, the full coordinate ascent scheme has updates as follows:

> perform one or more updates of $(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$ within an iterative scheme for the optimization problem:
>
> $$\underset{\boldsymbol{\mu}'_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}'_{q(\boldsymbol{\beta},\boldsymbol{u})}}{\text{argmax}} \left\{ \log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}'_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}'_{q(\boldsymbol{\beta},\boldsymbol{u})})^{[\boldsymbol{\beta},\boldsymbol{u}]} \right\}$$
>
> $B_{q(a)} \leftarrow \mu_{q(1/\sigma^2)} + A^{-2} \quad ; \quad \mu_{q(1/a)} \leftarrow 1/B_{q(a)}$
>
> $B_{q(\sigma^2)} \leftarrow \tfrac{1}{2}[\| \boldsymbol{\mu}_{q(\boldsymbol{u})} \|^2 + \text{tr}\{\boldsymbol{\Sigma}_{q(\boldsymbol{u})}\}] + \mu_{q(1/a)}$
>
> $\mu_{q(1/\sigma^2)} \leftarrow \tfrac{1}{2}(K+1)/B_{q(\sigma^2)}$.

For now, we deliberately leave the form of the $(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$ maximization strategy unspecified. We also allow for one or more updates of an iterative scheme aimed at maximizing $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}'_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}'_{q(\boldsymbol{\beta},\boldsymbol{u})})^{[\boldsymbol{\beta},\boldsymbol{u}]}$, rather than iterating to convergence at every iteration of the full coordinate ascent scheme. Section 3 describes various optimization strategies that could be used for updating $(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$.

Negligible absolute change in $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$ can be used as a stopping criterion for the iterations and an algebraic expression for this quantity is given in Appendix A.3.

We return to Example 2 in Section 4.3.

2.2.4 General Semiparametric Mean Field Variational Bayes Algorithm

We now treat semiparametric mean field variational Bayes in general, with the set-up laid out in Section 2 and restriction (8). Let $\log \underline{p}(\mathscr{D}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]}$ be defined with respect to the

factor graph of $p(\boldsymbol{x}, \boldsymbol{\phi}, \boldsymbol{\theta})$ with stochastic nodes $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ and $\boldsymbol{\phi}$. Algorithm 2 is a general coordinate ascent algorithm for approximate inference that builds on the standard mean field variational Bayes algorithm.

Initialize: $q(\boldsymbol{\theta}_2), \ldots, q(\boldsymbol{\theta}_M)$.

Cycle:

perform one or more updates of $\boldsymbol{\xi}$ within an iterative scheme for the optimization problem:

$$\underset{\boldsymbol{\xi}' \in \Xi}{\operatorname{argmax}} \left\{ \log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi}')^{[\phi]} \right\}$$

$$q(\boldsymbol{\theta}_1) \leftarrow \frac{\exp\left[E_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_1)\, q(\boldsymbol{\phi}; \boldsymbol{\xi})} \{\log\, p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})\}\right]}{\int \exp\left[E_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_1)\, q(\boldsymbol{\phi}; \boldsymbol{\xi})} \{\log\, p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})\}\right] d\boldsymbol{\theta}_1}$$

$$\vdots$$

$$q(\boldsymbol{\theta}_M) \leftarrow \frac{\exp\left[E_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_M)\, q(\boldsymbol{\phi}; \boldsymbol{\xi})} \{\log\, p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})\}\right]}{\int \exp\left[E_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_M)\, q(\boldsymbol{\phi}; \boldsymbol{\xi})} \{\log\, p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})\}\right] d\boldsymbol{\theta}_M}$$

until the absolute change in $\log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})$ is negligible.

Algorithm 2: *The general semiparametric mean field variational Bayes algorithm for restriction (8) with* $\log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})^{[\phi]}$ *defined with respect to factor graph of* $p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi})$ *with stochastic nodes* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ *and* $\boldsymbol{\phi}$.

For each of these approaches, there remains the practical problem of devising an optimization scheme for $\log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})^{[\phi]}$ and ensuring that it leads to the optimal parameters being chosen. Section 3 deals with this problem.

2.3 Relationship to Existing Literature

The general principle of semiparametric mean field variational Bayes that we have laid out in this section is not brand new and, in fact, instances of this principle have made appearances in the literature since the late 1990s — although they are few in number. We now briefly survey articles known to us that have a semiparametric mean field variational Bayes component. As we will see, the terminology varies quite considerably.

Barber and Bishop (1997) uses the terms *ensemble learning* and *hyperparameter adaptation* for what essentially is a semiparametric mean field variational Bayes approach to fitting multi-layer neural networks. They pre-specify Multivariate Normal distributions for the coefficient vector but, in their Section 2.1, allow covariance matrix parameters to be unspecified except for mean field assumptions. However, they do not include numerical details for minimizing Kullback-Leibler divergence over the Multivariate Normal parameters.

Honkela et al. (2010) adopt the phrase *fixed-form variational Bayes* in what is a quite general approach to semiparametric mean field variational Bayes as summarized in their Algorithm 1. Optimization of the pre-specified parametric component parameters is achieved using the *nonlinear conjugate gradient method*, which we describe in Section 3.2. However, Honkela et al. (2010) work with with *Riemannian* gradients which, they argue, are more efficient than Euclidean gradients.

In Knowles and Minka (2011) the focus is incorporation of pre-specified exponential family distributions whilst preserving the message passing aspect of a modular approach to mean field variational Bayes, known as *variational message passing*. They arrive at an extension which they label *non-conjugate variational message passing*. The exponential family distribution parameters are chosen via fixed-point iteration, which we describe in detail in Section 3.1.

Tan and Nott (2013) take a semiparametric mean field variational Bayes approach to approximate inference in Bayesian generalized linear mixed models for grouped data. They use pre-specified Multivariate Normal density functions for the random effects of each group with mean field product restrictions and achieve good approximation accuracy via so-called partially noncentered parameterizations.

In the case of pre-specified Multivariate Normal density functions, Wand (2014) obtains an explicit form for the fixed-point iteration scheme of Knowles and Minka (2011) and illustrated its use for the Poisson mixed model described in Section 2.2.3. In Luts and Wand (2015) and Menictas and Wand (2015), semiparametric mean field variational Bayes with Multivariate Normal pre-specification is applied, respectively, to count response semiparametric regression and heteroscedastic semiparametric regression.

2.4 Advantages and Limitations of Algorithm 2

As just described in Section 2.3, Algorithm 2 is a very useful generalization of the ordinary mean field variational Bayes algorithm and allows for tractable handling of a wider class of models. For example, the heteroscedastic nonparametric regression model of Menictas and Wand (2015) is such that ordinary mean field variational Bayes is numerically challenging if one uses the same product restrictions as used in homoscedastic nonparametric regression. The semiparametric mean field variational Bayes extension, based on Multivariate Normal pre-specification of basis function coefficients, leads to an iterative scheme with closed form updates. Simulation studies show very good accuracy compared with Markov chain Monte Carlo-based inference. However Algorithm 2 is not guaranteed to converge and, when it does converge, may result in mediocre approximate Bayesian inference. We close this section by briefly discussing such limitations of Algorithm 2.

In cases where a generic iterative procedure is used to solve $\text{argmax}_{\xi' \in \Xi}\{\log \underline{p}(\mathcal{D}; q, \xi')\}$ there is no guarantee that the lower bound is increased in a particular iteration or of convergence in general. As a consequence, the convergence guarantees enjoyed by ordinary mean field variational Bayes algorithms are not shared by their semiparametric extension. As we demonstrate in Section 4.2, convergence does not occur for particular numerical optimization strategies.

Lastly, there is the limitation imposed by the mean field restriction. Even though mean field variational Bayes can lead to very accurate approximate inference, there are

circumstances where its accuracy is quite poor. Some examples, with explanations for the inaccuracy, are given in Wang and Titterington (2005) and Neville et al. (2014). Semiparametric mean field variational Bayes shares this limitation since parametric pre-specification imposes a degradation in accuracy compared with ordinary mean field variational Bayes.

3. Numerical Optimization Strategies

In ordinary mean field variational Bayes, parameter optimization is achieved using a convex optimization algorithm that converges under reasonable assumptions (e.g. Luenberger and Ye, 2008). In the semiparametric extension there is no such convex optimization theory and general numerical optimization has to be called upon to optimize the parameters in the pre-specified parametric density function.

Numerical optimization is a major area of mathematical study with an enormous literature. Recent summaries are given in, for example, Givens and Hoeting (2005) and Ackleh et al. (2010), with the former being geared towards optimization problems arising in Statistics. The choice of optimization method typically is driven by factors such as the smoothness of the function requiring optimization and availability of expressions for low-order derivatives. Optimization methods with derivative information invariably take the form of iterative schemes. Semiparametric mean field variational Bayes often has the luxuries of smoothness and derivative expressions. It is also beneficial to have relatively simple iterative updates given the overarching goal of fast approximate inference. Therefore, we gear our summary of numerical optimization strategies towards simple derivative-based schemes. This is in keeping with the existing literature on parametric and semiparametric variational inference.

Let $f : D \subseteq \mathbb{R}^d \to \mathbb{R}$ be a function and consider the problem of finding its maximum or minimum value over a set $D_0 \subseteq D$. If all partial derivatives of f exist then a necessary condition for a maximum or minimum in the interior of D_0 is the *stationary point condition*

$$\frac{\partial}{\partial x_j} f(\boldsymbol{x}) = 0, \quad 1 \leq j \leq d. \tag{30}$$

This converts the optimization problem to a multivariate root-finding problem. However (30) is not a sufficient condition for global optima since local optima and saddle points of f inside D_0 also satisfy (30). Properties of f, such as regions where it is concave or convex, can aid the determination of global optima.

Throughout this section we make use of the derivative matrix and Hessian matrix notation defined in Appendix A.2.

3.1 Fixed-Point Iteration

Assuming that the derivative vector $\mathsf{D}_{\boldsymbol{x}} f(\boldsymbol{x})$ (defined formally in Appendix A.1) exists, the stationary point condition (30) can be written

$$\mathsf{D}_{\boldsymbol{x}} f(\boldsymbol{x})^T = \mathbf{0} \tag{31}$$

where $\mathbf{0}$ is the vector of zeroes. Fixed-point iteration aims to find points that satisfy (31), which we denote generically by $\boldsymbol{x}^*$. Such points are then candidates as maxima or minima

of f. Firstly, (31) is rewritten in the form

$$\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{x}) \tag{32}$$

for some function $\boldsymbol{g} : D \subseteq \mathbb{R}^d \to \mathbb{R}^d$. Given this $\boldsymbol{g}$, fixed-point iteration simply involves repeated evaluation of $\boldsymbol{g}$, as given in Algorithm 3

Initialize: $\boldsymbol{x} \leftarrow \boldsymbol{x}_{\text{init}}$ for some $\boldsymbol{x}_{\text{init}} \in D$.

Cycle:

$\boldsymbol{x} \leftarrow \boldsymbol{g}(\boldsymbol{x})$

until convergence.

Algorithm 3: *The fixed-point iteration algorithm in generic form.*

Note, however, the following issues regarding fixed point iteration:

- For a given stationary point condition (31) there are numerous functions $\boldsymbol{g}$ for which (32) holds. In other words, there are many possible fixed point algorithms available to solve (31).
- Once $\boldsymbol{g}$ and $\boldsymbol{x}_{\text{init}}$ are chosen then the above algorithm is *not necessarily guaranteed* to converge to a stationary point $\boldsymbol{x}^*$. There is a large literature on convergence of fixed-point iterative algorithms and good references on the topic include Section 8.1 of Ortega (1990) and Section 8.2 of Ackleh et al. (2010). For example Theorem 8.1.7 of Ortega (1990) asserts that convergence of Algorithm 3 is guaranteed when $\boldsymbol{x}_{\text{init}}$ is sufficiently close to $\boldsymbol{x}^*$, the components of $\boldsymbol{g}$ are differentiable at $\boldsymbol{x}^*$ and

  $$\rho\big(\mathsf{D}_{\boldsymbol{x}}\, \boldsymbol{g}(\boldsymbol{x}^*)\big) < 1.$$

  Here $\rho(\boldsymbol{A})$ denotes the *spectral radius* of the square matrix $\boldsymbol{A}$, defined to be

  $$\rho(\boldsymbol{A}) \equiv \text{maximum of the absolute values of the eigenvalues of } \boldsymbol{A}.$$

  Theorem 8.4 of Ackleh et al. (2010) provides a similar condition in terms of the *spectral norm* $\|\mathsf{D}_{\boldsymbol{x}}\, \boldsymbol{g}(\boldsymbol{x}^*)\|_{\text{spec}}$ where

  $$\|\boldsymbol{A}\|_{\text{spec}} \equiv \sqrt{\text{largest eigenvalue of } \boldsymbol{A}^T\boldsymbol{A}}.$$

- There are also theorems that guarantee convergence of Algorithm 3 for particular choices of $\boldsymbol{x}_{\text{init}}$. If D_0 is a closed convex subset of D such that $\boldsymbol{g} : D_0 \to D_0$, the entries of $\mathsf{D}_{\boldsymbol{x}}\, \boldsymbol{g}(\boldsymbol{x})$ are each bounded and continuous on D_0 and

  $$\sup_{\boldsymbol{x} \in D_0} \|\mathsf{D}_{\boldsymbol{x}}\, \boldsymbol{g}(\boldsymbol{x})\|_{\text{spec}} \leq \alpha < 1$$

  then Algorithm 3 will converge from any initial point $\boldsymbol{x}_{\text{init}} \in D_0$ (Theorems 8.2 and 8.3 of Ackleh *et al.*, 2010).

Despite this elegant theory, it is difficult to apply in practice with regards to choosing $\boldsymbol{g}$ and $\boldsymbol{x}_{\text{init}}$ so that Algorithm 3 converges. This is exemplified in Section 4.2 when we return to the Example 1 optimization problem. We also note that $\|\mathsf{D}_{\boldsymbol{x}}\,\boldsymbol{g}(\boldsymbol{x})\|_{\text{spec}} < 1$ near $\boldsymbol{x}^*$ is a *sufficient* but not *necessary* condition for convergence of fixed point iteration. Nevertheless, the function

$$\rho\big(\mathsf{D}_{\boldsymbol{x}}\,\boldsymbol{g}(\boldsymbol{x})\big)$$

is a useful convergence diagnostic for fixed-point iteration. For instance, if $\rho\big(\mathsf{D}_{\boldsymbol{x}}\,\boldsymbol{g}(\boldsymbol{x})\big) \gg 1$ during the iterations then this would indicate convergence problems and the possibility of non-existence of a fixed point $\boldsymbol{x}^*$.

Various adjustments to fixed-point iteration have been proposed to enhance convergence. For example, in the context of semiparametric mean field variational Bayes, Section 7 of Minka & Knowles (2011) describes a *damping* adjustment.

3.1.1 Newton-Raphson Iteration

Newton-Raphson iteration is a special case of fixed-point iteration for optimizing $\boldsymbol{f}$ with the $\boldsymbol{g}$ function taking the form

$$\boldsymbol{g}_{\text{NR}}(\boldsymbol{x}) = \boldsymbol{x} - \{\mathsf{H}_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})\}^{-1}\mathsf{D}_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})^T \tag{33}$$

where $\mathsf{H}_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})$ denotes the Hessian matrix of $\boldsymbol{f}(\boldsymbol{x})$ as formally defined in Appendix A.1. Assuming existence of $\{\mathsf{H}_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})\}^{-1}$, it is easily shown that $\boldsymbol{x} = \boldsymbol{g}_{\text{NR}}(\boldsymbol{x})$ if and only if $\mathsf{D}_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})^T = \boldsymbol{0}$. This leads to Algorithm 4, which conveys the generic form of Newton-Raphson iteration.

Initialize: $\boldsymbol{x} \leftarrow \boldsymbol{x}_{\text{init}}$ for some $\boldsymbol{x}_{\text{init}} \in D$.

Cycle:

$$\boldsymbol{x} \leftarrow \boldsymbol{x} - \{\mathsf{H}_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})\}^{-1}\mathsf{D}_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})^T$$

until convergence.

Algorithm 4: *The Newton-Raphson algorithm in generic form.*

Some pertinent features of Algorithm 4 are:

- The function $\boldsymbol{g}_{\text{NR}}$ in (33) has the property

  $$\rho\big(\mathsf{D}_{\boldsymbol{x}}\,\boldsymbol{g}_{\text{NR}}(\boldsymbol{x}^*)\big) = 0 \tag{34}$$

  for stationary points $\boldsymbol{x}^*$, A proof is given in Appendix A.4. Therefore, via Theorem 8.4 of Ackleh et al. (2010), convergence to $\boldsymbol{x}^*$ is guaranteed from a sufficiently close $\boldsymbol{x}_{\text{init}}$.

- If $\boldsymbol{x}_{\text{init}}$ is such that Algorithm 4 is convergent to $\boldsymbol{x}^*$ then, under certain regularity conditions, convergence is *quadratic*, in that the number of significant figures doubles on each iteration.

- Locating $\boldsymbol{x}_{\text{init}}$ values sufficiently close to $\boldsymbol{x}^*$ for convergence to occur can be difficult in practice and it is common to combine Newton-Raphson iteration with more robust optimization strategies, such as the Nelder-Mead simplex method.

- A disadvantage of Newton-Raphson iteration compared with other fixed-point iterative schemes is the requirement for second order partial derivatives, corresponding to the entries of the Hessian matrix $\mathsf{H}_{\boldsymbol{x}} f(\boldsymbol{x})$. A feeling for the type of additional calculus needed is given in Appendix A.7.

- A variant of Newton-Raphson optimization known as *damped* Newton-Raphson employs line searches (or backtracking) in order to achieve much improved convergence behavior. See, e.g., Section 9.5.2 of Boyd and Vandenberghe (2004).

3.2 Nonlinear Conjugate Gradient Method

The *nonlinear conjugate gradient method* is based on the *conjugate gradient method*, an established iterative approach to solving large systems of linear equations (Hestenes and Stiefel, 1952). The former arises from applying the latter to the linear system that arises from optimization of a multivariate quadratic function. Details of the nonlinear conjugate gradient method are given in Section 10.8 of Press et al. (2007). Algorithm 5 lists the *Polak-Ribiére* version of the nonlinear conjugate gradient method for *maximization* of f over D. Since minimization of f is equivalent to maximization of $-f$ it is straightforward to adapt Algorithm 5 to the minimization problem. We choose the Polak-Ribiére form here, but another one is the *Fletcher-Reeves* form given by $\beta \leftarrow (\boldsymbol{v}_{\text{curr}}^T \boldsymbol{v}_{\text{curr}})/(\boldsymbol{v}_{\text{prev}}^T \boldsymbol{v}_{\text{prev}})$.

Initialize: $\boldsymbol{x} \leftarrow \boldsymbol{x}_{\text{init}}$ for some $\boldsymbol{x}_{\text{init}} \in D$.

$\boldsymbol{v}_{\text{prev}} \leftarrow \mathsf{D}_{\boldsymbol{x}} f(\boldsymbol{x})^T \;\; ; \;\; \alpha \leftarrow \underset{\alpha>0}{\text{argmax}}\, f(\boldsymbol{x} + \alpha\, \boldsymbol{v}_{\text{prev}})$

$\boldsymbol{x} \leftarrow \boldsymbol{x} + \alpha\, \boldsymbol{v}_{\text{prev}} \;\; ; \;\; \boldsymbol{s} \leftarrow \boldsymbol{v}_{\text{prev}}$

Cycle:

$\boldsymbol{v}_{\text{curr}} \leftarrow \mathsf{D}_{\boldsymbol{x}} f(\boldsymbol{x})^T \;\; ; \;\; \beta \leftarrow \boldsymbol{v}_{\text{curr}}^T(\boldsymbol{v}_{\text{curr}} - \boldsymbol{v}_{\text{prev}})/(\boldsymbol{v}_{\text{prev}}^T \boldsymbol{v}_{\text{prev}})$

$\boldsymbol{s} \leftarrow \beta\, \boldsymbol{s} + \boldsymbol{v}_{\text{curr}} \;\; ; \;\; \alpha \leftarrow \underset{\alpha>0}{\text{argmax}}\, f(\boldsymbol{x} + \alpha\, \boldsymbol{s})$

$\boldsymbol{x} \leftarrow \boldsymbol{x} + \alpha\, \boldsymbol{s} \;\; ; \;\; \boldsymbol{v}_{\text{prev}} \leftarrow \boldsymbol{v}_{\text{curr}}$

until convergence.

Algorithm 5: *The nonlinear conjugate gradient method for maximization of the function f with the Polak-Ribiére form of the β parameter.*

A key aspect of the nonlinear conjugate gradient method is that, on each iteration, it takes a step in the direction $\mathsf{D}\, f(\boldsymbol{x})^T$ from the current position at $\boldsymbol{x}$. This is the steepest instantaneous direction on the f surface. The parameter denoted by β has several alter-

native forms. Nonlinear conjugate gradient methods have been shown to have good global convergence properties (Dai and Yuan, 1999).

3.3 Other Optimization Strategies

Other popular optimization strategies include *ascent (or descent) algorithms* (e.g. Boyd and Vandenberghe, 2004, Section 9.3), *quasi-Newton methods* (e.g. Givens and Hoeting, 2005, Section 2.2.2.3), the *Gauss-Newton method* (e.g. Givens and Hoeting, 2005, Section 2.2.3), *stochastic gradient descent* (e.g. Bottou, 2004) and the *Nelder-Mead simplex method* (Nelder and Mead, 1965). The last of these has the attraction of not requiring derivatives of f and is generally more robust than derivative-based methods.

3.4 Application to Semiparametric Mean Field Variational Bayes

We now focus on the optimization component of Algorithm 2

$$\underset{\boldsymbol{\xi}' \in \Xi}{\operatorname{argmax}} \left\{ \log \underline{p}(\mathscr{D}; q, \boldsymbol{\xi}')^{[\boldsymbol{\phi}]} \right\} \tag{35}$$

and discuss ways in which numerical optimization strategies described in Sections 3.1–3.3 are applicable.

The stationary condition for the maximizer in (35) is

$$\mathsf{D}_{\boldsymbol{\xi}} \log \underline{p}(\mathscr{D}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]^T} = \mathbf{0}$$

and this may be manipulated in any of a number of ways to produce an equation of the form $\boldsymbol{\xi} = \boldsymbol{g}(\boldsymbol{\xi})$ for some function $\boldsymbol{g}$. Fixed-point iteration Algorithm 3 can then be entertained but, as discussed in Section 3.1, converge is not guaranteed for arbitrary $\boldsymbol{g}$. We study this issue in the context of Examples 1 and 2 in Sections 4.2 and 4.3.

Newton-Raphson iteration involves iterative updates:

$$\boldsymbol{\xi} \leftarrow \boldsymbol{\xi} - \{\mathsf{H}_{\boldsymbol{\xi}} \log \underline{p}(\mathscr{D}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]}\}^{-1} \mathsf{D}_{\boldsymbol{\xi}} \log \underline{p}(\mathscr{D}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]^T}$$

and so is a candidate for insertion into Algorithm 2 for updating the pre-specified parametric q-density parameters.

Another alternative is, of course, updating $\boldsymbol{\xi}$ according to one or more iterations of the nonlinear conjugate gradient method given by Algorithm 5, or any other iterative optimization scheme. However, convergence needs to be monitored. For high-dimensional $\boldsymbol{\xi}$, *speed* of convergence may be also be an important factor. Next we discuss an adjustment aimed at improving the convergence speed of gradient-based algorithms.

3.4.1 Riemannian Geometry Adjustment

As explained in, for example, Section 6.2 of Murray and Rice (1993) the density function family $\{q(\boldsymbol{\phi}; \boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi\}$ can be viewed as a *submanifold* of a *Riemannian manifold*. Riemannian manifolds do not necessarily have a *flat* Euclidean geometry. For example, the Riemannian manifold corresponding to the Univariate Normal family:

$$\left\{ \frac{1}{\sqrt{2\pi\sigma^2_{q(\phi)}}} \exp\left\{ -\frac{(\phi - \mu_{q(\phi)})^2}{2\sigma^2_{q(\phi)}} \right\} : \mu_{q(\phi)} \in \mathbb{R}, \sigma^2_{q(\phi)} > 0 \right\} \tag{36}$$

has *hyperbolic geometry* (Murray and Rice, 1993, Example 6.6.2) which is *curved*. Therefore notions such as closeness of two members of (36) and steepness of gradients when searching over the parameter space $\Xi = \mathbb{R} \times \mathbb{R}_+$ are not properly captured by the Euclidean geometry notions of distance and slope. Adjustments for the Riemannian geometry of the family often improve convergence of optimization algorithms for solving problems such as (35). More detailed discussion on this issue is given in Section 2.2 of Honkela et al. (2010) and Section 2.3 of Hoffman et al. (2013).

Consider an optimization method that uses gradients of the form

$$\mathsf{D}_{\boldsymbol{\xi}} \log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]^T}$$

to find the direction of steepest descent of the objective function $\log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]}$. The Riemannian geometry adjustment is to instead use

$$[-E\{\mathsf{H}_{\boldsymbol{\xi}} \log q(\boldsymbol{\phi}; \boldsymbol{\xi})\}]^{-1} \mathsf{D}_{\boldsymbol{\xi}} \log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]^T} \tag{37}$$

where the premultiplying matrix is the inverse *Fisher information* of $q(\boldsymbol{\phi}; \boldsymbol{\xi})$. In the Machine Learning literature (e.g. Amari, 1998) (37) is often labeled the *natural* or *Riemannian gradient* of $\log \underline{p}(\boldsymbol{\mathscr{D}}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]}$ with respect to $\boldsymbol{\xi}$ and the corresponding geometry is called *information geometry*. If $q(\phi; \boldsymbol{\xi})$ corresponds to the Univariate Normal family (36) then the Fisher information matrix is $\text{diag}(\sigma^{-2}_{q(\phi)}, \frac{1}{2}\sigma^{-4}_{q(\phi)})$. Therefore, from (37), the natural gradient of $\log \underline{p}(q; \mu_{q(\phi)}, \sigma^2_{q(\phi)})^{[\phi]}$ with respect to $(\mu_{q(\phi)}, \sigma^2_{q(\phi)})$ is given by

$$\begin{bmatrix} \sigma^2_{q(\phi)} \dfrac{\partial \underline{p}(q; \mu_{q(\phi)}, \sigma^2_{q(\phi)})^{[\phi]}}{\partial \mu_{q(\phi)}} \\ 2\sigma^4_{q(\phi)} \dfrac{\partial \underline{p}(q; \mu_{q(\phi)}, \sigma^2_{q(\phi)})^{[\phi]}}{\partial \sigma^2_{q(\phi)}} \end{bmatrix}^T . \tag{38}$$

Honkela et al. (2010) is a major contribution to semiparametric mean field variational Bayes methodology and their Algorithm 1 uses the nonlinear conjugate gradient method (Algorithm 5) with natural gradients rather than ordinary gradients. Via both simple examples and numerical studies, they make a compelling case for the use of natural gradients for optimization of the parameters of the pre-specified parametric q-density function.

3.5 Summary of Semiparametric Mean Field Variational Bayes Ramifications

In this section we have discussed several iterative numerical optimization strategies. Any of these are candidates for the updating $\boldsymbol{\xi}$ in the Algorithm 2 cycle. Special mention has been given to the well-known Newton-Raphson iteration since it can achieve very rapid convergence and the nonlinear conjugate gradient method which has been shown to be effective in semiparametric mean field variational Bayes contexts when the Riemmannian geometry adjustment of Section 3.4.1 is employed (Honkela et al., 2010).

General fixed-point iteration has been discussed in detail. It has the advantage of yielding particularly simple iterative updates for $\boldsymbol{\xi}$ in Algorithm 2. Established theory shows that the spectral radius of the derivative matrix of the fixed-point function can be

used to assess the quality of the scheme. In Section 4 we will explain how a particular fixed-point iteration scheme, which we call *natural* fixed-point iteration, has attractive properties when $q(\cdot;\boldsymbol{\xi})$ is an exponential family density function. We will also revisit Examples 1 and 2 in Section 4 and make some comparisons among various numerical optimization strategies. Natural fixed-point iteration is seen to perform particularly well.

4. Exponential Family Special Case

We now focus on the important special case where the parametric density function family $\{q(\boldsymbol{\phi};\boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi\}$ can be expressed in exponential family form:

$$q(\boldsymbol{\phi};\boldsymbol{\eta}) = \exp\{\boldsymbol{T}(\boldsymbol{\phi})^T\boldsymbol{\eta} - A(\boldsymbol{\eta})\}\, h(\boldsymbol{\phi}), \quad \boldsymbol{\eta} \in H, \tag{39}$$

where $\boldsymbol{\eta}$ is a one-to-one transformation of $\boldsymbol{\xi}$ and H is the image of Ξ under this transformation. In (39) $A(\boldsymbol{\eta})$ is called the *log-partition function* and $h(\boldsymbol{\phi})$ is called the *base measure*. For example, the Univariate Normal density function family used in Example 1:

$$q(\phi;\mu_{q(\phi)},\sigma^2_{q(\phi)}) = \frac{1}{\sqrt{2\pi\sigma^2_{q(\phi)}}}\,\exp\left\{\frac{-(\phi-\mu_{q(\phi)})^2}{2\sigma^2_{q(\phi)}}\right\}, \quad \mu_{q(\phi)} \in \mathbb{R}, \sigma^2_{q(\phi)} > 0,$$

can be expressed as (39) with

$$\boldsymbol{T}(\phi) = \begin{bmatrix} \phi \\ \phi^2 \end{bmatrix}, \quad \boldsymbol{\eta} \equiv \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \mu_{q(\phi)}/\sigma^2_{q(\phi)} \\ -1/(2\sigma^2_{q(\phi)}) \end{bmatrix}, \quad A(\boldsymbol{\eta}) = -\tfrac{1}{4}\,\eta_1^2/\eta_2 - \tfrac{1}{2}\log(-2\eta_2)$$

and $h(\phi) = (2\pi)^{-1/2}$. The natural parameter space is $H = \{(\eta_1,\eta_2) : \eta_1 \in \mathbb{R}, \eta_2 < 0\}$. Even though semiparametric mean field variational Bayes can involve pre-specification of an arbitrary parametric family, virtually all methodology and examples in the existing literature involves pre-specification within an exponential family. Exponential family distributions also play an important role in the theory of mean field variational Bayes (e.g. Sato, 2001; Beal and Ghahramani, 2006; Wainwright and Jordan, 2008).

Now consider the general factor graph set-up described in Section 2.2.1 with the approximate marginal log-likelihood $\log \underline{p}(\boldsymbol{\mathscr{D}};q,\boldsymbol{\eta})^{[\boldsymbol{\phi}]}$ given by (13) but as a function of the natural parameter vector $\boldsymbol{\eta}$. Define

$$\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\eta})\} \equiv \sum_{j\in\,\text{neighbors}(\boldsymbol{\phi})} E_{q(\phi;\boldsymbol{\eta})}\{\log(f_j)\}$$

so that

$$\log \underline{p}(\boldsymbol{\mathscr{D}};q,\boldsymbol{\eta})^{[\boldsymbol{\phi}]} = \text{Entropy}\{q(\boldsymbol{\phi};\boldsymbol{\eta})\} + \text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\eta})\}.$$

An advantage of working with exponential family density functions is that the entropy takes the simple form:

$$\text{Entropy}\{q(\boldsymbol{\phi};\boldsymbol{\eta})\} = A(\boldsymbol{\eta}) - \mathsf{D}_{\boldsymbol{\eta}}\,A(\boldsymbol{\eta})\,\boldsymbol{\eta} + E[\exp\{h(\boldsymbol{\phi})\}]$$

where the notation of Appendix A.2 is being used. Moreover, as shown in Lemma 1 of Appendix A.5, the derivative vector of $\text{Entropy}\{q(\boldsymbol{\phi};\boldsymbol{\eta})\}$ is simply

$$\mathsf{D}_{\boldsymbol{\eta}}\,\text{Entropy}\{q(\boldsymbol{\phi};\boldsymbol{\eta})\} = -\boldsymbol{\eta}^T\,\mathsf{H}_{\boldsymbol{\eta}}\,A(\boldsymbol{\eta}). \tag{40}$$

This implies that the stationary point condition

$$\{\mathsf{D}_{\boldsymbol{\eta}} \log \underline{p}(q;\boldsymbol{\eta})^{[\boldsymbol{\phi}]}\}^T = \mathbf{0} \tag{41}$$

is equivalent to

$$\boldsymbol{\eta} = \{\mathsf{H}_{\boldsymbol{\eta}}\, A(\boldsymbol{\eta})\}^{-1} \mathsf{D}_{\boldsymbol{\eta}} \text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\eta})\}^T. \tag{42}$$

Algorithm 1 of Knowles and Minka (2011) is a fixed-point iteration scheme based on (42).

One further interesting and useful connection concerns the *mean parameter* vector

$$\boldsymbol{\tau} \equiv E\{T(\boldsymbol{\phi})\}$$

which is related to the natural parameter vector via

$$\boldsymbol{\tau} = \mathsf{D}_{\boldsymbol{\eta}}\, A(\boldsymbol{\eta})^T.$$

Under suitable technical conditions $\boldsymbol{\tau}$ is a one-to-one transformation of $\boldsymbol{\eta}$. Also the chain rule for differentiation of a smooth function s, listed as Lemma 2 in Appendix A.5, is

$$\mathsf{D}_{\boldsymbol{\eta}}\, s = (\mathsf{D}_{\boldsymbol{\tau}}\, s)(\mathsf{D}_{\boldsymbol{\eta}}\boldsymbol{\tau}) = (\mathsf{D}_{\boldsymbol{\tau}}\, s)\, D_{\boldsymbol{\eta}}\{\mathsf{D}_{\boldsymbol{\eta}}\, A(\boldsymbol{\eta})^T\} = (\mathsf{D}_{\boldsymbol{\tau}}\, s)\, \mathsf{H}_{\boldsymbol{\eta}}\, A(\boldsymbol{\eta}).$$

which leads to

$$\mathsf{D}_{\boldsymbol{\tau}}\, s = \{\mathsf{H}_{\boldsymbol{\eta}}\, A(\boldsymbol{\eta})\}^{-1} \mathsf{D}_{\boldsymbol{\eta}}\, s. \tag{43}$$

Putting all of these relationships together we have:

Result 1 *Let $\boldsymbol{\xi}$ be an arbitrary differentiable one-to-one transformation of $\boldsymbol{\eta}$. Then the stationary point condition (41) is equivalent to each the following conditions:*

(a) $\boldsymbol{\eta} = \{\mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})\}^{-1}\, \mathsf{D}_{\boldsymbol{\eta}} \text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\eta})\}^T$,

(b) $\boldsymbol{\eta} = \{\mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})\}^{-1} (\mathsf{D}_{\boldsymbol{\eta}}\, \boldsymbol{\xi})^T\, \mathsf{D}_{\boldsymbol{\xi}} \text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\xi})\}^T$,

(c) $\boldsymbol{\eta} = \mathsf{D}_{\boldsymbol{\tau}}\, \text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\tau})\}^T$ *and*

(d) $\boldsymbol{\eta} = (\mathsf{D}_{\boldsymbol{\tau}}\, \boldsymbol{\xi})^T\, \mathsf{D}_{\boldsymbol{\xi}} \text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\xi})\}^T$.

We make the following remarks concerning Result 1:

- Result 1(a) immediately gives rise to the following fixed-point iteration scheme in the natural parameter space $\boldsymbol{\eta} \in H$:

$$\boldsymbol{\eta} \leftarrow \{\mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})\}^{-1} \mathsf{D}_{\boldsymbol{\eta}} \text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\eta})\}^T. \tag{44}$$

  We refer to (44) as the *natural fixed-point iteration scheme* and denote the corresponding fixed-point function by

$$\boldsymbol{g}_{\text{nat}}(\boldsymbol{\eta}) \equiv \{\mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})\}^{-1} \mathsf{D}_{\boldsymbol{\eta}} \text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\eta})\}^T.$$

  According to the theory of fixed-point iteration discussed in Section 3.1, convergence of (44) is implied by

$$\rho\big(\mathsf{D}_{\boldsymbol{\eta}}\, \boldsymbol{g}_{\text{nat}}(\boldsymbol{\eta})\big) < 1$$

  in a neighborhood of the maximizer $\boldsymbol{\eta}^*$.

- Result 1(b)–(d) offer the possibility of more convenient forms for the fixed-point updates in terms of derivatives of the common parameters or mean parameters. Particularly noteworthy is the fact that Result 1(c)–(d) do not require computation of $\{\mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})\}^{-1}$. We make use of this situation for the Multivariate Normal family in Section 4.1.

- The Fisher information of $q(\boldsymbol{\phi} : \boldsymbol{\eta})$ is

$$-E\{\mathsf{H}_{\boldsymbol{\eta}} \log\, q(\boldsymbol{\phi}; \boldsymbol{\eta})\} = \mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})$$

  which implies that the natural fixed-point iteration scheme (44) involves updating $\boldsymbol{\eta}$ according to natural Riemannian gradients of NonEntropy$(q; \boldsymbol{\tau})$. From Result 1(c), an equivalent updating scheme is

$$\boldsymbol{\eta} \leftarrow \mathsf{D}_{\boldsymbol{\tau}} \text{NonEntropy}\{q(\boldsymbol{\phi}; \boldsymbol{\tau})\}^T$$

  which simply involves updating $\boldsymbol{\eta}$ according to the direction of maximum slope on the NonEntropy$(q; \boldsymbol{\tau})$ surface in the $\boldsymbol{\tau}$ space.

- The forms for the stationary point in Result 1 can also be used to derive iterative Newton-Raphson schemes for maximizing $\log \underline{p}(q; \boldsymbol{\eta})^{[\boldsymbol{\phi}]}$. An example, corresponding to Result 1(a) and optimization within the $\boldsymbol{\eta}$ space, is

$$\begin{aligned}\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} - \left[\mathsf{H}_{\boldsymbol{\eta}} \text{NonEntropy}\{q(\boldsymbol{\phi}; \boldsymbol{\eta})\} - \mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) - (\boldsymbol{\eta}^T \otimes \boldsymbol{I}) \mathsf{D}_{\boldsymbol{\eta}} \text{vec}\{\mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})\}\right]^{-1} \\ \times [\mathsf{D}_{\boldsymbol{\eta}} \text{NonEntropy}\{q(\boldsymbol{\phi}; \boldsymbol{\eta})\}^T - \mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) \boldsymbol{\eta}].\end{aligned}$$

The vec operator flattens a square matrix into a column vector and is defined formally in Appendix A.1.

Any of the other optimization methods mentioned in Section 3 can also be applied to the problem of obtaining

$$\boldsymbol{\eta}^* \equiv \underset{\boldsymbol{\eta} \in H}{\text{argmax}} \{\log \underline{p}(\mathscr{D}; q, \boldsymbol{\xi})^{[\boldsymbol{\phi}]}\} = \underset{\boldsymbol{\eta} \in H}{\text{argmax}} \left[A(\boldsymbol{\eta}) - \mathsf{D}_{\boldsymbol{\eta}}\, A(\boldsymbol{\eta}) \boldsymbol{\eta} + \text{NonEntropy}\{q(\boldsymbol{\phi}; \boldsymbol{\eta})\}\right]$$

and those involving gradients benefit from the entropy derivative result (40). Additionally, relationship (43) implies that natural (Riemannian) gradients of the objective function in the natural parameter space are equivalent to ordinary Euclidean gradients in the mean parameter space.

4.1 Multivariate Normal Special Case

We now focus on the important special case of $q(\boldsymbol{\phi}; \boldsymbol{\xi})$ being a d-variate Multivariate Normal density function:

$$q(\boldsymbol{\phi}; \boldsymbol{\mu}_{q(\boldsymbol{\phi})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}|^{-1/2} \exp\{-\tfrac{1}{2} (\boldsymbol{\phi} - \boldsymbol{\mu}_{q(\boldsymbol{\phi})})^T \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}^{-1} (\boldsymbol{\phi} - \boldsymbol{\mu}_{q(\boldsymbol{\phi})})\}.$$

Let

$$\boldsymbol{\xi} \equiv \begin{bmatrix} \boldsymbol{\mu}_{q(\boldsymbol{\phi})} \\ \text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}) \end{bmatrix}$$

be the vector of common parameters. An explicit form for the natural fixed point iteration updates in terms of $\boldsymbol{\mu}_{q(\boldsymbol{\phi})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}$ was derived by Wand (2014) and appears as equation (7) there. However Result 1 affords a more direct derivation of the same result, that benefits from (43) and the cancellation of the $\mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})$ matrix. We can also obtain a neater alternative explicit form by using a differentiation identity, given as Lemma 4 in Appendix A.5. The essence of Lemma 4 is given in the appendix of Opper and Archambeau (2009).

Result 2 *Natural fixed-point iteration for $q(\boldsymbol{\phi};\boldsymbol{\xi})$ corresponding to the $N(\boldsymbol{\mu}_{q(\boldsymbol{\phi})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})})$ density function is equivalent to the following updating scheme:*

$$\left\{\begin{array}{rcl} \boldsymbol{v}_{q(\boldsymbol{\phi})} & \leftarrow & \mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\phi})}} \text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\phi})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})})^T \\ \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} & \leftarrow & -\{\mathsf{H}_{\boldsymbol{\mu}_{q(\boldsymbol{\phi})}} \text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\phi})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})})\}^{-1} \\ \mu_{q(\boldsymbol{\phi})} & \leftarrow & \boldsymbol{\mu}_{q(\boldsymbol{\phi})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} \boldsymbol{v}_{q(\boldsymbol{\phi})}. \end{array}\right.$$

Appendix A.6 provides details on how Result 2 follows from Result 1.

Result 2 facilitates a semiparametric mean field variational Bayes algorithm that requires only the first and second order derivatives of $\text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\phi})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})})$ with respect to $\boldsymbol{\mu}_{q(\boldsymbol{\phi})}$. Concrete illustrations are given in Section 4.3 and Appendix A of Menictas and Wand (2015).

In the case where $q(\phi;\boldsymbol{\xi})$ is the Univariate Normal density function with mean $\mu_{q(\phi)}$ and variance $\sigma^2_{q(\phi)}$ Result 2 leads to the following common parameter updates for the natural fixed point iterative scheme:

$$\left\{\begin{array}{rcl} v_{q(\phi)} & \leftarrow & \dfrac{\partial \,\text{NonEntropy}(q; \mu_{q(\phi)}, \sigma^2_{q(\phi)})}{\partial \mu_{q(\phi)}} \\ \sigma^2_{q(\phi)} & \leftarrow & -1 \Big/ \left\{ \dfrac{\partial^2 \,\text{NonEntropy}(q; \mu_{q(\phi)}, \sigma^2_{q(\phi)})}{\partial \mu^2_{q(\phi)}} \right\} \\ \mu_{q(\boldsymbol{\phi})} & \leftarrow & \mu_{q(\boldsymbol{\phi})} + \sigma^2_{q(\phi)}\, v_{q(\phi)}. \end{array}\right. \tag{45}$$

Despite its use of natural gradients, there is no automatic guarantee that iteration of (45) leads to convergence to the maximum of $\log \underline{p}(\mathscr{D}; q, \boldsymbol{\xi})^{[\phi]}$ on any given cycle of Algorithm 2. However, the fixed-point iteration theory summarized in Section 3.1 provides some guidance. We now use Example 1 to illustrate this point using the natural fixed-point iteration scheme (45), an alternative simpler fixed-point scheme and a Newton-Raphson scheme.

4.2 Application to Example 1

Consider again the Gumbel random sample example introduced in Section 2.2.2 and the problem of minimum Kullback-Leibler approximation of $p(\phi|\boldsymbol{x})$ within the Univariate Normal family. As shown there, the optimization problem is encapsulated in (19) and (20).

The Newton-Raphson scheme that arises directly from (19) is

$$\begin{bmatrix} \mu_{q(\phi)} \\ \sigma^2_{q(\phi)} \end{bmatrix} \leftarrow \begin{bmatrix} \mu_{q(\phi)} \\ \sigma^2_{q(\phi)} \end{bmatrix} - \{\mathsf{H} f^N_{\text{Ex1}}(\mu_{q(\phi)}, \sigma^2_{q(\phi)}; n, \textstyle\sum_{i=1}^n e^{-x_i}, \mu_\phi, \sigma^2_\phi)\}^{-1} \\ \times \mathsf{D}\, f^N_{\text{Ex1}}(\mu_{q(\phi)}, \sigma^2_{q(\phi)}; n, \textstyle\sum_{i=1}^n e^{-x_i}, \mu_\phi, \sigma^2_\phi)^T. \tag{46}$$

Differentiation with respect to $(\mu_\phi, \sigma_\phi^2)$ is suppressed in the D and H on the right-hand side of (46). Simple calculus leads to (46) being equivalent to the fixed-point iterative scheme

$$\begin{bmatrix} \mu_{q(\phi)} \\ \sigma^2_{q(\phi)} \end{bmatrix} \leftarrow \boldsymbol{g}_{\text{NR}} \left(\begin{bmatrix} \mu_{q(\phi)} \\ \sigma^2_{q(\phi)} \end{bmatrix}; n, \sum_{i=1}^n e^{-x_i}, \mu_\phi, \sigma_\phi^2 \right) \tag{47}$$

where

$$\begin{aligned} \boldsymbol{g}_{\text{NR}} \left(\begin{bmatrix} x \\ y \end{bmatrix}; a, b, c, d \right) \quad \equiv \quad & \begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} -b\, e^{x+\frac{1}{2}y} - 1/d & -\frac{1}{2}\, b\, e^{x+\frac{1}{2}y} \\ -\frac{1}{2}\, b\, e^{x+\frac{1}{2}y} & -\frac{1}{2y^2} - \frac{1}{4}\, b\, e^{x+\frac{1}{2}y} \end{bmatrix}^{-1} \\ & \times \begin{bmatrix} a - b\, e^{x+\frac{1}{2}y} - (x-c)/d \\ \frac{1}{2y} - \frac{1}{2}\, b e^{x+\frac{1}{2}y} - 1/(2d) \end{bmatrix}. \end{aligned}$$

According to (45), the natural fixed-point iteration scheme (44) takes the form (47), but with $\boldsymbol{g}_{\text{NR}}$ replaced by $\boldsymbol{g}_{\text{nat}}$ where

$$\boldsymbol{g}_{\text{nat}} \left(\begin{bmatrix} x \\ y \end{bmatrix}; a, b, c, d \right) \equiv \begin{bmatrix} x + \{a - b\, e^{x+\frac{1}{2}y} - (x-c)/d\}/(b\, e^{x+\frac{1}{2}y} + d^{-1}) \\ 1/(b\, e^{x+\frac{1}{2}y} + d^{-1}) \end{bmatrix}.$$

Lastly, there is the very simple fixed-point iteration scheme that arises from full simplification of $\mathsf{D}\, f^N_{\text{Ex1}}(\mu_{q(\phi)}, \sigma^2_{q(\phi)}; n, \sum_{i=1}^n e^{-x_i}, \mu_\phi, \sigma_\phi^2)^T = \mathbf{0}$, and corresponds to fixed points of

$$\boldsymbol{g}_{\text{simp}} \left(\begin{bmatrix} x \\ y \end{bmatrix}; a, b, c, d \right) \equiv \begin{bmatrix} c + d(a - b\, e^{x+\frac{1}{2}y}) \\ 1/(b\, e^{x+\frac{1}{2}y} + d^{-1}) \end{bmatrix}.$$

In Figure 4 we compare $\boldsymbol{g}_{\text{NR}}$, $\boldsymbol{g}_{\text{nat}}$ and $\boldsymbol{g}_{\text{simp}}$ in terms of the behavior of the spectral norm function $\rho\big(\mathsf{D}\, \boldsymbol{g}(x,y)\big)$ and convergence of the fixed-point iterative scheme. We simulated data from the $n = 20$ version of the Gumbel random sample model (14) with the value of ϕ set to 0. The sufficient statistic $\sum_{i=1}^{20} \exp(-x_i)$ fully determines f^N_{Ex1} and has a mean of 20. In an effort to exhibit typical behavior, we selected a sample that produced a sufficient statistic value close to this mean. The actual value is $\sum_{i=1}^{20} \exp(-x_i) \approx 19.94$. The hyperparameters were set to $\mu_\phi = 0$ and $\sigma_\phi^2 = 10^{10}$. The optimal parameters in the minimum Kullback-Leibler Univariate Normal approximation to $p(\phi|\boldsymbol{x})$ are $(\mu^*_{q(\phi)}, (\sigma^2_{q(\phi)})^*) \approx (0.2260, 0.0500)$. We set up a 101×101 pixel mesh of $(\mu_{q(\phi)}, \log(\sigma^2_{q(\phi)}))$ values around this optimum with limits $(\mu^*_{q(\phi)} - 5, \mu^*_{q(\phi)} + 5)$ and $\big(\log\{(\sigma^*_{q(\phi)}/5)^2\}, \log\{(5\sigma^*_{q(\phi)})^2\}\big)$. The upper panels of Figure 4 show the

$$\text{indicator of } \ \rho\big(\mathsf{D}\boldsymbol{g}(\mu_{q(\phi)}, \sigma^2_{q(\phi)})\big) < 1 \ \text{ for } \boldsymbol{g} \in \{\boldsymbol{g}_{\text{NR}}, \boldsymbol{g}_{\text{nat}}, \boldsymbol{g}_{\text{simp}}\}.$$

The lower panels show the

$$\text{indicator of fixed-point iteration converging when starting from } (\mu_{q(\phi)}, \sigma^2_{q(\phi)}).$$

The top half of Figure 4 shows, via dark grey shading, that both $\rho\big(\mathsf{D}\, \boldsymbol{g}_{\text{NR}}(\mu_{q(\phi)}, \sigma^2_q)\big)$ and $\rho\big(\mathsf{D}\, \boldsymbol{g}_{\text{nat}}(\mu_{q(\phi)}, \sigma^2_q)\big)$ are below 1 in regions around the root. The dark grey region for

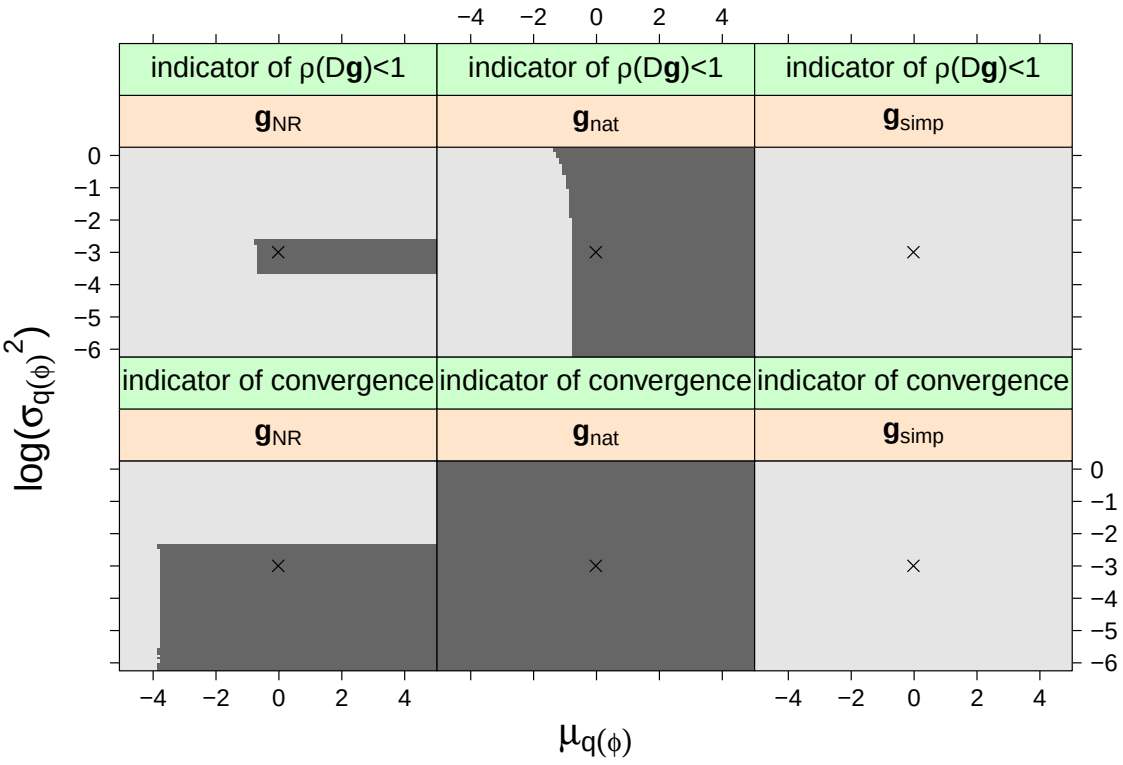

Figure 4: *Upper panels: Dark grey shading shows points where* $\rho\big(\mathrm{D}\,\boldsymbol{g}(\mu_{q(\phi)}, \sigma_q^2)\big) < 1$ *for* $\boldsymbol{g} \in \{\boldsymbol{g}_{\mathrm{NR}}, \boldsymbol{g}_{\mathrm{nat}}, \boldsymbol{g}_{simp}\}$. *Lower panels: Dark grey shading shows points from which fixed-point iteration, based on* $\boldsymbol{g} \in \{\boldsymbol{g}_{\mathrm{NR}}, \boldsymbol{g}_{\mathrm{nat}}, \boldsymbol{g}_{simp}\}$, *converges if initialized from that point. The optimum is shown by a cross in each panel and corresponds to minimum Kullback-Leibler divergence for a single* $n = 20$ *sample of the Gumbel random sample model with hyperparameters set to* $\mu_\phi = 0$ *and* $\sigma_\phi^2 = 10^{10}$.

$\boldsymbol{g}_{\mathrm{nat}}$ is much larger than that of $\boldsymbol{g}_{\mathrm{NR}}$, suggesting that the former has better convergence properties according to the theory described in Section 3.1. The lower panels confirm this, with $\boldsymbol{g}_{\mathrm{nat}}$-based fixed-point iteration converging from every initial value on the pixel mesh, but Newton-Raphson iteration not converging from the sub-region on the top and left-hand side of the mesh. Also note that $\rho\big(\mathrm{D}\,\boldsymbol{g}_{\mathrm{simp}}(\mu_{q(\phi)}, \sigma_q^2)\big) \geq 1$ over the whole pixel mesh and $\boldsymbol{g}_{\mathrm{simp}}$-based fixed-point iteration does not converge, regardless of initial point.

Figure 5 shows the iteration trajectories from four different starting values and four iterative schemes based on the same data and hyperparameter values as used in Figure 4. Also shown in each panel is an image plot of the surface being maximized, with a logarithmic scale used for $\sigma^2_{q(\phi)}$. In addition to the fixed-point iteration schemes based on $\boldsymbol{g}_{\mathrm{NR}}$ and $\boldsymbol{g}_{\mathrm{nat}}$ we include the nonlinear conjugate gradient method given in Algorithm 5 and

the Riemannian geometry adjustment involving the natural gradients given by (38). In most cases the iterations led to convergence to $(\mu^*_{q(\phi)}, (\sigma^2_{q(\phi)})^*)$ and the first three iterates are plotted. However, Newton-Raphson failed to converge from the starting values in each of the upper panels and the subsequent iterates are outside of the image plot boundaries.

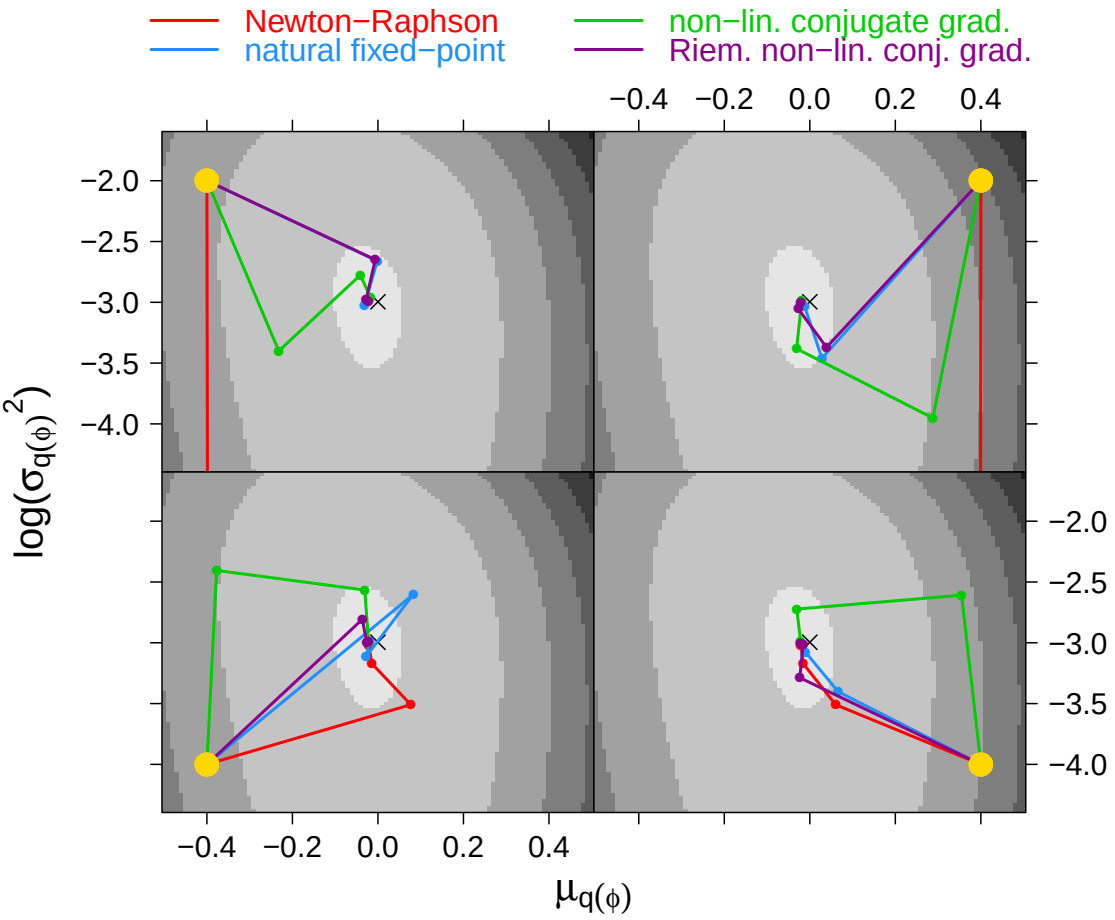

Figure 5: *Iteration trajectories of four iterative algorithms aimed at solving the minimum Kullback-Leibler problem for a Gumbel random sample of size $n = 20$ with hyperparameters $\mu_{q(\phi)} = 0$ and $\sigma^2_{q(\phi)} = 10^{10}$. The initial value differs for each panel and is shown by the yellow dot. The iterative algorithms are: (1) Newton-Raphson fixed-point iteration based on $\boldsymbol{g}_{\text{NR}}$, (2) natural fixed-point iteration based on $\boldsymbol{g}_{\text{nat}}$, (3) ordinary non-linear conjugate gradient method and (4) Riemannian non-linear conjugate gradient method.*

The most striking feature of Figure 5 is the directness with which natural fixed-point iteration and the Riemannian non-linear conjugate gradient method converge from all four starting points and the similarity of their trajectories. This behavior is in keeping with the fact that both work with the more appropriate Riemannian gradients. The ordinary non-linear conjugate gradient trajectories are not as direct. Similar observations are made in Honkela et al. (2010). As demonstrated there, the payoffs from using Riemannian gradients in non-linear conjugate gradient updating are greater in higher-dimensional versions of semiparametric mean field variational Bayes. Based on Figure 5, we anticipate that natural fixed-point iteration is also very good in higher dimensions, and this is corroborated by experiments for Example 2 described in Section 4.3. Newton-Raphson fixed-point iteration is seen to be unreliable for this optimization problem and nowhere near as robust as natural fixed-point iteration. Lastly, we note that the behavior represented in Figures 4 and 5 persists across each of several other samples that we generated.

4.3 Application to Example 2

From (29) and some simple matrix algebra

$$\begin{aligned}
&\text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \\
&\quad = \boldsymbol{y}^T \boldsymbol{C} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} - \mathbf{1}^T \exp\left\{ \boldsymbol{C} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2} \text{diagonal}(\boldsymbol{C} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{C}^T) \right\} \\
&\qquad -\tfrac{1}{2} \text{tr}\left(\begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)} \boldsymbol{I}_K \end{bmatrix} \{ \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^T + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \} \right) \\
&\qquad -\tfrac{1}{2}(p+K)\log(2\pi) - \tfrac{1}{2}\, p \log(\sigma_{\boldsymbol{\beta}}^2) - \tfrac{1}{2}\, K\, E_q\{\log(\sigma^2)\} - \mathbf{1}^T \log(\boldsymbol{y}!)
\end{aligned}$$

where

$$\mu_{q(1/\sigma^2)} = E_{q(1/\sigma^2)}(1/\sigma^2).$$

The derivatives appearing in Result 2 are

$$\begin{aligned}
\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}} \text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})^T &= \boldsymbol{C}^T \left[\boldsymbol{y} - \exp\{ \boldsymbol{C} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2} \text{diagonal}(\boldsymbol{C} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{C}^T) \} \right] \\
&\quad - \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)} \boldsymbol{I}_K \end{bmatrix} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}
\end{aligned}$$

and

$$\begin{aligned}
&\mathsf{H}_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}} \text{NonEntropy}(q; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}) = \\
&-\left(\boldsymbol{C}^T \text{diag}[\exp\{ \boldsymbol{C} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2} \text{diagonal}(\boldsymbol{C} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{C}^T) \}]\, \boldsymbol{C} + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)} \boldsymbol{I}_K \end{bmatrix} \right).
\end{aligned}$$

It follows that the updates take the explicit form

$$\left\{ \begin{aligned}
\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\leftarrow \exp\{ \boldsymbol{C} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2} \text{diagonal}(\boldsymbol{C} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{C}^T) \} \\
\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\leftarrow \left(\boldsymbol{C}^T \text{diag}\{ \boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} \} \boldsymbol{C} + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)} \boldsymbol{I}_K \end{bmatrix} \right)^{-1} \\
\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} &\leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \left\{ \boldsymbol{C}^T (\boldsymbol{y} - \boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}) - \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2} \boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)} \boldsymbol{I}_K \end{bmatrix} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \right\}.
\end{aligned} \right.$$

This is equivalent to the fixed-point iteration scheme

$$\begin{bmatrix} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \\ \text{vech}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \end{bmatrix} \leftarrow \boldsymbol{g}_{\text{Ex2}}\left(\begin{bmatrix} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \\ \text{vech}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \end{bmatrix} \; ; \; \boldsymbol{y}, \boldsymbol{C}, \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma^2)}\boldsymbol{I}_K \end{bmatrix} \right)$$

where

$$\boldsymbol{g}_{\text{Ex2}}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \text{vech}(\boldsymbol{\Sigma}) \end{bmatrix} \; ; \; \boldsymbol{y}, \boldsymbol{C}, \boldsymbol{M} \right) \equiv \begin{bmatrix} \boldsymbol{\mu} + \big[\boldsymbol{C}^T \text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{C} + \boldsymbol{M}\big]^{-1} \\ \times \big[\boldsymbol{C}^T\{\boldsymbol{y} - \boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\} - \boldsymbol{M}\boldsymbol{\mu}\big] \\ \text{vech}\Big(\big[\boldsymbol{C}^T \text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{C} + \boldsymbol{M}\big]^{-1}\Big) \end{bmatrix}$$

and

$$\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv \exp\{\boldsymbol{C}\boldsymbol{\mu} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T)\}.$$

Note that the vech operator stores the unique entries of a symmetric matrix in a column vector. A formal definition of vech is given in Appendix A.1.

We simulated data according to the following special case of the Poisson mixed model:

$$\begin{gathered} y_{ij}|U_i \sim \text{Poisson}\left\{\exp(\beta_0 + \beta_1\, x_{ij} + U_i)\right\},\; U_i|\,\sigma^2 \sim N(0,\sigma^2), \\ 1 \le i \le m,\; 1 \le j \le n, \quad \boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_{\boldsymbol{\beta}}^2\,\boldsymbol{I}), \\ \sigma^2|\,a \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a), \quad a \sim \text{Inverse-Gamma}(\tfrac{1}{2}, A^{-2}). \end{gathered} \tag{48}$$

The hyperparameters were set at $\sigma_{\boldsymbol{\beta}} = A = 10^5$ and the sample sizes were $m = 30$, $n = 5$. Note that (48) is a special case of (21) with $\boldsymbol{Z} = \boldsymbol{I}_m \otimes \boldsymbol{1}_n$, where $\boldsymbol{1}_n$ is the $n \times 1$ vector with all entries equal to one. We then ran Algorithm 2 with $q(\boldsymbol{\beta},\boldsymbol{u})$ pre-specified to be the $N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$ density function and a single natural fixed-point iteration in each cycle based on $\boldsymbol{g}_{\text{Ex2}}$. The fixed-point iteration search over values of $[\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^T \text{ vech}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})]^T$ is within an open subset of $\mathbb{R}^{560}$.

Figure 6 shows trace plots of $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$ and $\rho\big(\mathsf{D}\boldsymbol{g}_{\text{Ex2}}\big)$, based on the explicit expressions for $\mathsf{D}\boldsymbol{g}_{\text{Ex2}}$ given in Appendix A.7. The upper panel indicates that the algorithm becomes close to convergence after $6 - 10$ iterations. After the same number of iterations the values of $\rho\big(\mathsf{D}\boldsymbol{g}_{\text{Ex2}}\big)$ fall below 1 and settle at about 0.15.

Before leaving this example we note that numerical checks indicate that the optimal $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ matrix is approximately sparse, with dominant diagonal entries. This implies the possibility of low-rank approximations to the above semiparametric mean field variational Bayes algorithm given, as described in Section 4.1.3 of Challis and Barber (2013).

5. A Non-Exponential Family Example

In the previous section it was seen that semiparametric mean field variational Bayes with $q(\boldsymbol{\phi};\boldsymbol{\xi})$ having an exponential family form (39) leads to simplifications of the Kullback-Leibler minimization problem. However, $q(\boldsymbol{\phi};\boldsymbol{\xi})$ does not have to be restricted in this way. In this section we illustrate semiparametric mean field variational Bayes with the q-density of $\boldsymbol{\phi}$ specified to be in a non-exponential family: the family of Skew-Normal density functions (Azzalini and Dalla Valle, 1996). Even though this family has a multivariate extension,

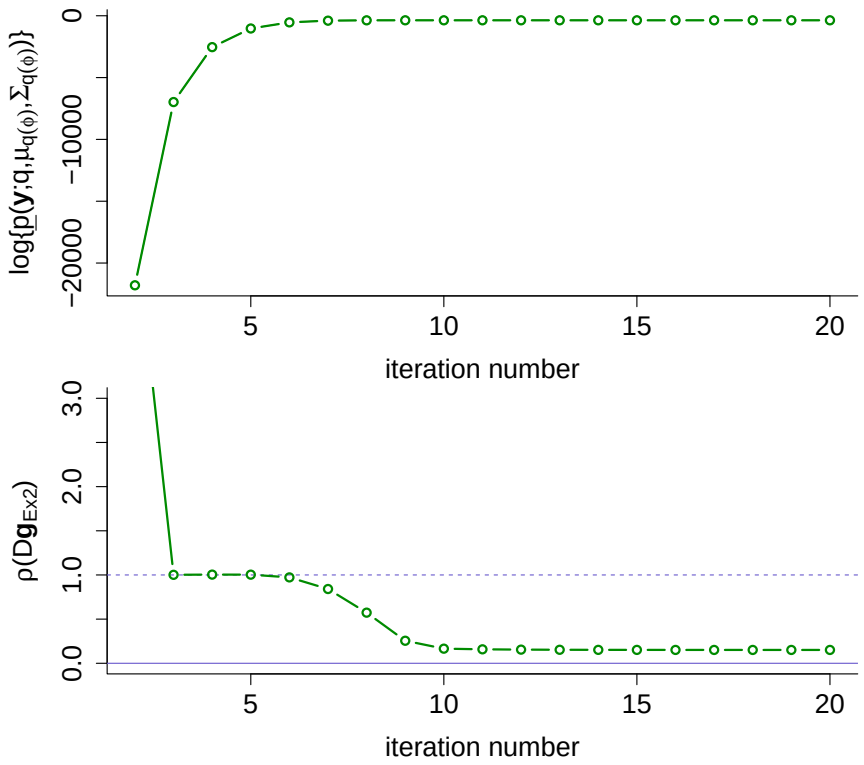

Figure 6: *Trace plots of* $\log \underline{p}(\boldsymbol{y}; q, \boldsymbol{\xi})^{[(\boldsymbol{\beta},\boldsymbol{u})]}$ *and* $\rho\big(\mathsf{D}\boldsymbol{g}_{\mathrm{Ex2}}\big)$ *for the version of the Poisson mixed model given by (48) with sample sizes* $m = 30$ *and* $n = 5$.

we restrict attention to the univariate case and, in particular, its use within the context of Example 1.

Specification of $q(\phi; \boldsymbol{\xi})$ being within the family of univariate Skew-Normal density functions entails having

$$q(\phi; \boldsymbol{\xi}) = \sqrt{\frac{2}{\pi \sigma^2_{q(\phi)}}} \exp\left\{ -\frac{(\phi - \mu_{q(\phi)})^2}{2\sigma^2_{q(\phi)}} \right\} \Phi\left\{ \frac{\lambda_{q(\phi)}(\phi - \mu_{q(\phi)})}{\sigma_{q(\phi)}} \right\} \tag{49}$$

where $\Phi(x) \equiv (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-t^2/2}\, dt$ is the $N(0,1)$ cumulative distribution function. The q-density parameter vector is $\boldsymbol{\xi} = (\mu_{q(\phi)}, \sigma^2_{q(\phi)}, \lambda_{q(\phi)})$ and the corresponding parameter space is $\boldsymbol{\Xi} = \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}$. Now consider the Example 1 setting with $q(\phi; \boldsymbol{\xi})$ restricted to the Skew-Normal family (49). The marginal log-likelihood lower bound, given by (17) and (18),

depends on the explicit expressions

$$E_{q(\phi;\boldsymbol{\xi})}(\phi) \;=\; \mu_{q(\phi)} + \frac{\sigma_{q(\phi)}\lambda_{q(\phi)}}{\sqrt{(\pi/2)(1+\lambda^2_{q(\phi)})}}, \quad \mathrm{Var}_{q(\phi;\boldsymbol{\xi})}(\phi) = \sigma^2_{q(\phi)}\left\{1 - \frac{2\lambda^2_{q(\phi)}}{\pi(1+\lambda^2_{q(\phi)})}\right\}$$

$$\text{and} \quad M_{q(\phi;\boldsymbol{\xi})}(t) \;=\; 2\exp\left(\mu_{q(\phi)} + \tfrac{1}{2}\sigma^2_{q(\phi)}t^2\right)\Phi\left(\frac{\lambda_{q(\phi)}\sigma_{q(\phi)}t}{\sqrt{1+\lambda^2_{q(\phi)}}}\right)$$

where, as defined in Section 2.2.2, $M_{q(\phi;\boldsymbol{\xi})}$ is the moment generating function corresponding to $q(\phi;\boldsymbol{\xi})$. It also depends on

$$\text{Entropy}\{q(\phi;\boldsymbol{\xi})\} = \tfrac{1}{2}\{1 + \log(\pi/2) + \log(\sigma^2_{q(\phi)})\} - \sqrt{\frac{2}{\pi}}\int_{-\infty}^{\infty}\log\Phi(\lambda_{q(\phi)}t)\Phi(\lambda_{q(\phi)}t)e^{-t^2/2}\,dt$$

which does not simplify any further. Plugging these expressions into (17) and (18) we get the Kullback-Leibler optimal Skew-Normal q-density function is $q(\phi;\mu^*_{q(\phi)},(\sigma^2_{q(\phi)})^*,\lambda^*_{q(\phi)})$ where

$$\begin{bmatrix} \mu^*_{q(\phi)} \\ (\sigma^2_{q(\phi)})^* \\ \lambda^*_{q(\phi)} \end{bmatrix} = \underset{\mu_{q(\phi)}\in\mathbb{R},\sigma^2_{q(\phi)}>0,\lambda_{q(\phi)}\in\mathbb{R}}{\text{argmax}} \left\{ f^{\text{SN}}_{\text{Ex1}}\left(\mu_{q(\phi)},\sigma^2_{q(\phi)},\lambda_{q(\phi)};n,\sum_{i=1}^n e^{-x_i},\mu_\phi,\sigma^2_\phi\right)\right\} \quad (50)$$

and

$$\begin{aligned} f^{\text{SN}}_{\text{Ex1}}(x,y,z;a,b,c,d) \;=\;& \tfrac{1}{2}\log(y) - \sqrt{\frac{2}{\pi}}\int_{-\infty}^{\infty}\log\{\Phi(zt)\}\Phi(zt)e^{-t^2/2}\,dt \\ &+a\left\{x + z\sqrt{\frac{2y}{\pi(1+z^2)}}\right\} - 2b\,\exp(x+\tfrac{1}{2}y)\Phi\left(\frac{z\sqrt{y}}{\sqrt{z^2+1}}\right) \\ &-\frac{1}{2d}\left[\left\{x + z\sqrt{\frac{2y}{\pi(1+z^2)}} - c\right\}^2 + y\left\{1 - \frac{2z^2}{\pi(1+z^2)}\right\}\right]. \end{aligned}$$

Optimization problem (50) is considerably more challenging than its Normal counterpart. In particular, evaluations of the objective function and its derivatives require numerical integration.

We solved (50) for three Gumbel random samples of size $n = 5, 10$ and 20 and with $\sum_{i=1}^n e^{-x_i} \approx n$, corresponding to the mean of this sufficient statistic. The intractable integral in $f^{\text{SN}}_{\text{Ex1}}$ was approximated using a trapezoidal quadrature scheme similar to that described in Appendix B.2 of Wand et al. (2011). The limits of the trapezoidal grid were increased until the ratio of the global maximum and minimum absolute values of the integrand fell below 10^{-20}. The number of grid points was then doubled until the relative difference between two successive iterations was less than 10^{-20}. Multiple start locations and simulated annealing were used to locate global optima. Natural fixed-point iteration no

longer applies in this non-exponential family example and optimization of $f_{\text{Ex1}}^{\text{SN}}$ was accomplished using the *Broyden-Fletcher-Goldfarb-Shanno* quasi-Newton method via the `optim()` function in the R computing environment (R Development Core Team, 2016).

Figure 7 shows the optimal Skew-Normal q-density functions, together with the exact posterior density functions and those based on the Normal q-density restriction. We see that the Normal approximation is inferior for very low sample sizes, but that the approximations are about the same for moderate to large sample sizes.

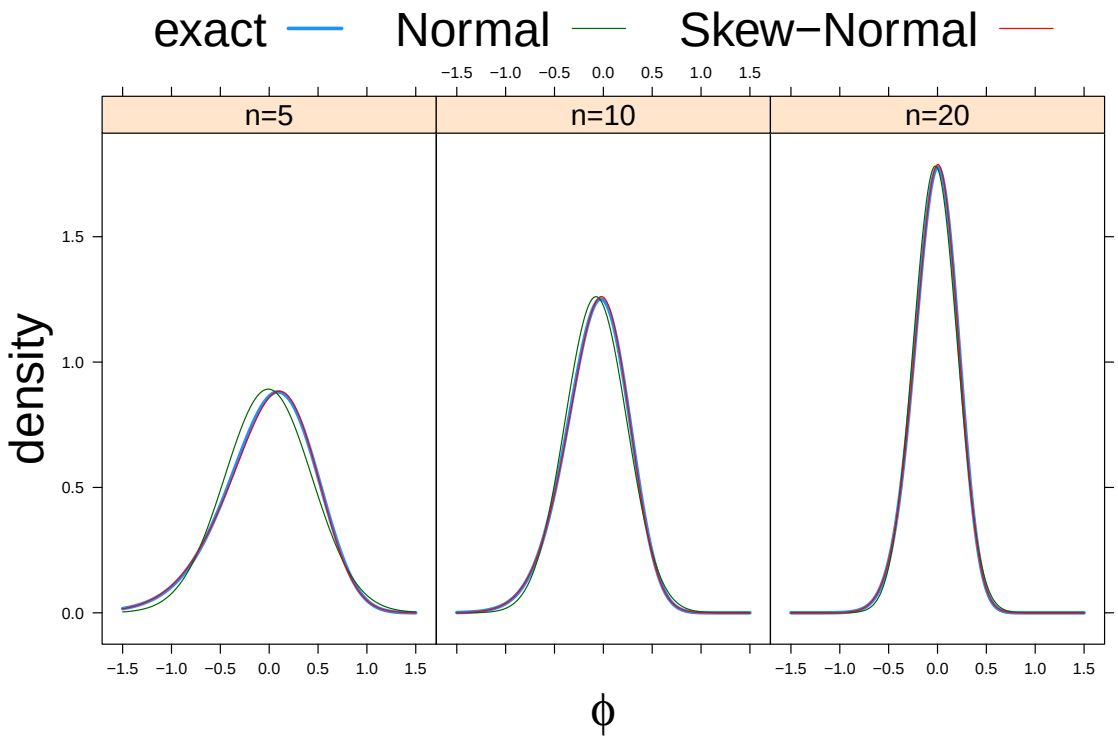

Figure 7: *Skew-Normal minimum Kullback-Leibler approximate posterior density functions for samples of size* $n = 5, 10$ *and* 20 *for the Example 1 Gumbel random sample setting. The exact posterior density functions and those based on restriction to the Normal family are also shown.*

6. Closing Remarks

We have taken a broad view of mean field variational Bayes with parametric pre-specification of one of the q-density components and coined the term 'semiparametric mean field variational Bayes' for this general approach. As well as laying out the general principles of

semiparametric mean field variational Bayes, we have provided an overview of the numerical issues attached to this methodology. Natural fixed-point iteration has been identified as a promising general approach to dealing with the Kullback-Leibler optimization problem and its attractive Riemannian gradient properties have been elucidated. Proof of convergence of a particular semiparametric mean field variational Bayes strategy appears to be too difficult a goal. However, for fixed-point iteration strategies, the spectral radius of the derivative matrix of the fixed-point update function is a reasonable diagnostic measure for checking convergence.

Acknowledgments

This research was partially supported by Australian Research Council Discovery Project DP110100061. We are grateful to Frances Kuo, Christian Wolff and Rob Womersley for their advice on aspects of this research. We also thank the editor and referees for their helpful comments.

Appendix A. Definitions and Derivations

A.1 Matrix definitions and identity

If $\boldsymbol{A}$ is a $d \times d$ matrix then $\text{vec}(\boldsymbol{A})$ is the $d^2 \times 1$ vector obtained by stacking the columns of $\boldsymbol{A}$ underneath each other in order from left to right. The inverse vec operator is denoted by vec^{-1}. In addition we let $\text{vech}(\boldsymbol{A})$ denote the $\frac{1}{2}\,d(d+1) \times 1$ vector obtained from $\text{vec}(\boldsymbol{A})$ by eliminating the above-diagonal entries of $\boldsymbol{A}$. If $\boldsymbol{A}$ is symmetric then $\text{vech}(\boldsymbol{A})$ contains all of the unique entries of $\boldsymbol{A}$.

The derivations also require the commutation and duplication matrix notation of Magnus and Neudecker (1999). If $\boldsymbol{A}$ is an arbitrary $d \times d$ matrix then the *commutation matrix of order* d, denoted by $\boldsymbol{K}_d$, is the the $d^2 \times d^2$ matrix of zeroes and ones for which

$$\boldsymbol{K}_d \,\text{vec}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A}^T).$$

If $\boldsymbol{B}$ is a symmetric but otherwise arbitrary $d \times d$ matrix then the *duplication matrix of order* d is the $d^2 \times \frac{1}{2}\,d(d+1)$ matrix of zeroes and ones for which

$$\boldsymbol{D}_d \text{vech}(\boldsymbol{B}) = \text{vec}(\boldsymbol{B}).$$

The Moore-Penrose inverse of $\boldsymbol{D}_d$ is

$$\boldsymbol{D}_d^+ \equiv (\boldsymbol{D}_d^T \boldsymbol{D}_d)^{-1} \boldsymbol{D}_d^T.$$

Note that

$$\boldsymbol{D}_d^+ \text{vec}(\boldsymbol{B}) = \text{vech}(\boldsymbol{B}). \tag{51}$$

Another useful notation is

$$\boldsymbol{Q}(\boldsymbol{A}) \equiv (\boldsymbol{A} \otimes \boldsymbol{1}^T) \odot (\boldsymbol{1}^T \otimes \boldsymbol{A})$$

for a general $m \times n$ matrix $\boldsymbol{A}$ and $\boldsymbol{1}$ a $n \times 1$ vector of ones. The symbol $\odot$ denotes element-wise product.

The following well-known matrix identity is used several times in the derivations:

$$\text{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^T \otimes \boldsymbol{A})\,\text{vec}(\boldsymbol{B}). \tag{52}$$

A.2 Derivative Matrix and Hessian Matrix Notation

Our summary of derivative-based optimization, and subsequent discussion, benefits from derivative vector and Hessian matrix notation. Such notation is not universal, and throughout this article we follow the conventions of Magnus and Neudecker (1999).

If $\boldsymbol{h}$ is a $\mathbb{R}^p$-valued with argument $\boldsymbol{x} \in \mathbb{R}^d$ then the *derivative matrix* of $\boldsymbol{h}$ with respect to $\boldsymbol{x}$, denoted by $\mathsf{D}_{\boldsymbol{x}}\boldsymbol{h}(\boldsymbol{x})$, is the $p \times d$ matrix with (i, j) entry

$$\frac{\partial\, \boldsymbol{h}(\boldsymbol{x})_i}{\partial\, x_j}$$

A concrete derivative vector example is given in Section 2.3 of Wand (2014).

In the case $p = 1$, the *Hessian matrix* of $\boldsymbol{h}$ with respect to $\boldsymbol{x}$ is the $d \times d$ matrix

$$\mathsf{H}_{\boldsymbol{x}}\boldsymbol{h}(\boldsymbol{x}) \equiv \mathsf{D}_{\boldsymbol{x}}[\{\mathsf{D}_{\boldsymbol{x}}\,\boldsymbol{h}(\boldsymbol{x})\}^T].$$

A.3 Example 2 Derivational Details

Here provide derivational details pertaining to Example 2 discussed in Section 2.2.3.

According to product restriction (22), the optimal q-density functions satisfy

$$\begin{aligned} q^*(\boldsymbol{\beta}, \boldsymbol{u}) &\propto \exp[E_{q(\sigma^2,a)} \log\{p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, a)\}], \\ q^*(\sigma^2) &\propto \exp[E_{q(\boldsymbol{\beta},\boldsymbol{u},a)} \log\{p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, a)\}] \\ \text{and} \quad q^*(a) &\propto \exp[E_{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2)} \log\{p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, a)\}] \end{aligned}$$

(e.g. Bishop, 2006, Section 10.1.1). Simple algebraic steps lead to the forms given in (23).

First we consider general pre-specified q-density families of the form $q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi}),\ \boldsymbol{\xi} \in \Xi$. With the help of (10) each of the terms in (27), can be expressed as follows:

$$\begin{aligned}
\text{Entropy}\{q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})\} &= -\textstyle\int_{\mathbb{R}^{K+2}} \log\{q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})\} q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})\, d\boldsymbol{\beta}\, d\boldsymbol{u}, \\
\text{Entropy}\{q(\sigma^2)\} &= \log(B_{q(\sigma^2)}) + \tfrac{1}{2}(K+1) + \log\{\Gamma(\tfrac{1}{2}(K+1))\} \\
&\quad -\tfrac{1}{2}(K+3)\text{digamma}\{\tfrac{1}{2}(K+1)\}, \\
\text{Entropy}\{q(a)\} &= \log(B_{q(a)}) + 1 - 2\,\text{digamma}(1), \\
E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u})\} &= \boldsymbol{y}^T\{\boldsymbol{X}\, E_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{\beta}) + \boldsymbol{Z}\, E_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{u})\} \\
&\quad -\mathbf{1}^T E_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}\{\exp(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\} - \mathbf{1}^T \log(\boldsymbol{y}!), \\
E_q\{\log p(\boldsymbol{\beta}, \boldsymbol{u}|\, \sigma^2)\} &= -\tfrac{1}{2}(p+K)\log(2\pi) - \tfrac{1}{2}\, p \log(\sigma^2_{\boldsymbol{\beta}}) \\
&\quad -\tfrac{1}{2}\, K\, \{\log\{B_{q(\sigma^2)}\} - \text{digamma}\{\tfrac{1}{2}(K+1)\}\} \\
&\quad -\tfrac{1}{2\sigma^2_{\boldsymbol{\beta}}}\left[\| E_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{\beta})\|^2 + \text{tr}\{\text{Cov}_{q(\beta,\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{\beta})\}\right] \\
&\quad -\tfrac{1}{2}\, \mu_{q(1/\sigma^2)}\left[\| E_{q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{u})\|^2 + \text{tr}\{\text{Cov}_{q(\beta,\boldsymbol{u};\boldsymbol{\xi})}(\boldsymbol{u})\}\right], \\
E_q\{\log p(\sigma^2|\, a)\} &= -\tfrac{1}{2}\log(\pi) - \tfrac{1}{2}\,\{\log\{B_{q(a)}\} - \text{digamma}(1)\} \\
&\quad -\tfrac{3}{2}\,\{\log\{B_{q(\sigma^2)}\} - \text{digamma}\{\tfrac{1}{2}(K+1)\} \\
&\quad -\mu_{q(1/a)}\mu_{q(1/\sigma^2)} \\
\text{and}\quad E_q\{\log p(a)\} &= -\tfrac{1}{2}\log(\pi) - \log(A) - \tfrac{3}{2}\,\{\log\{B_{q(a)}\} - \text{digamma}(1)\} \\
&\quad -\mu_{q(1/a)}/A^2.
\end{aligned} \tag{53}$$

The $(\boldsymbol{\beta}, \boldsymbol{u})$-localized approximate marginal log-likelihood expression given by (28) follows immediately from the relevant terms in (53).

If $q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\xi})$ is specified to be the $N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$ density function then the terms in (27) that depend on $\boldsymbol{\xi} = (\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$ are

$$\begin{aligned}
\text{Entropy}\{q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})\} &= \tfrac{1}{2}(p+K)\{1 + \log(2\pi)\} + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}|, \\
E_q\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u})\} &= \boldsymbol{y}^T\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} - \mathbf{1}^T\log(\boldsymbol{y}!) \\
&\quad -\mathbf{1}^T\exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\} \\
\text{and}\quad E_q\{\log p(\boldsymbol{\beta}, \boldsymbol{u}|\, \sigma^2)\} &= -\tfrac{1}{2}(p+K)\log(2\pi) - \tfrac{1}{2}\, p\log(\sigma^2_{\boldsymbol{\beta}}) \\
&\quad -\tfrac{1}{2}\, K\,\{\log\{B_{q(\sigma^2)}\} - \text{digamma}\{\tfrac{1}{2}(K+1)\} \\
&\quad -\tfrac{1}{2\sigma^2_{\boldsymbol{\beta}}}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right\} \\
&\quad -\tfrac{1}{2}\, \mu_{q(1/\sigma^2)}\left\{\| \boldsymbol{\mu}_{q(\boldsymbol{\mu})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u})})\right\}
\end{aligned} \tag{54}$$

where $\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}]$. The $(\boldsymbol{\beta}, \boldsymbol{u})$-localized approximate marginal log-likelihood expression given by (29) follows immediately. An explicit expression for $\log \underline{p}(q; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$, for use a a stopping criterion, can be formed by combining the relevant terms from (53) with those in (54).

A.4 Proof of (34)

In this proof, all appearances of D and H are assumed to be with respect to $\boldsymbol{x}$. Let

$$\boldsymbol{g}_{\rm NR}(\boldsymbol{x}) \equiv \boldsymbol{x} - \{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}\, \mathsf{D}\, f(\boldsymbol{x})^T.$$

Then, using (52),

$$\begin{aligned} d\,\boldsymbol{g}_{\rm NR}(\boldsymbol{x}) &= d\boldsymbol{x} + \{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}\{d\mathsf{H}\, f(\boldsymbol{x})\}\{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}\mathsf{D}\, f(\boldsymbol{x})^T \\ &\quad -\{\mathsf{H}\, f(\boldsymbol{x})\}^{-1} d\,\mathsf{D}\, f(\boldsymbol{x})^T \\ &= d\boldsymbol{x} + \{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}\mathrm{vec}^{-1}[\mathsf{D}\mathrm{vec}\{\mathsf{H}\, f(\boldsymbol{x})\}\, d\boldsymbol{x}]\{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}\mathsf{D}\, f(\boldsymbol{x})^T \\ &\quad -\{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}\mathsf{H}\, f(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}\mathrm{vec}\Big(\boldsymbol{I}\,\mathrm{vec}^{-1}[\mathsf{D}\,\mathrm{vec}\{\mathsf{H}\, f(\boldsymbol{x})\}\, d\boldsymbol{x}]\{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}\mathsf{D}\, f(\boldsymbol{x})^T\Big) \\ &= \{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}\Big([\mathsf{D}\, f(\boldsymbol{x})\{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}] \otimes \boldsymbol{I}\Big)\mathsf{D}\,\mathrm{vec}\{\mathsf{H}\, f(\boldsymbol{x})\}\, d\boldsymbol{x}. \end{aligned}$$

Therefore

$$\mathsf{D}\,\boldsymbol{g}_{\rm NR}(\boldsymbol{x}) = \{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}\Big([\mathsf{D}\, f(\boldsymbol{x})\{\mathsf{H}\, f(\boldsymbol{x})\}^{-1}] \otimes \boldsymbol{I}\Big)\mathsf{D}\,\mathrm{vec}\{\mathsf{H}\, f(\boldsymbol{x})\}.$$

Since $\mathsf{D}\, f(\boldsymbol{x}^*) = \boldsymbol{0}$, we get

$$\mathsf{D}\,\boldsymbol{g}_{\rm NR}(\boldsymbol{x}^*) = \boldsymbol{O},$$

where $\boldsymbol{O}$ is the $d \times d$ matrix with all entries equal to zero, and (34) follows immediately.

A.5 Lemmas and Proofs Required for Results 1 and 2

Lemma 1 *If*

$$q(\boldsymbol{x};\boldsymbol{\eta}) = \exp\{\boldsymbol{T}(\boldsymbol{x})^T\boldsymbol{\eta} - A(\boldsymbol{\eta})\}h(\boldsymbol{x})$$

is an exponential family density function then

$$\mathsf{D}_{\boldsymbol{\eta}}\,\mathrm{Entropy}\{q(\boldsymbol{x};\boldsymbol{\eta})\} = -\boldsymbol{\eta}^T\,\mathsf{H}_{\boldsymbol{\eta}}\,A(\boldsymbol{\eta}).$$

Proof of Lemma 1

Since

$$\mathrm{Entropy}\{q(\boldsymbol{x};\boldsymbol{\eta})\} = A(\boldsymbol{\eta}) - \mathsf{D}_{\boldsymbol{\eta}}\,A(\boldsymbol{\eta})\,\boldsymbol{\eta} - E[\log\{h(\boldsymbol{x})\}]$$

we then have

$$\mathsf{D}_{\boldsymbol{\eta}}\,\mathrm{Entropy}\{q(\boldsymbol{x};\boldsymbol{\eta})\} = \mathsf{D}_{\boldsymbol{\eta}}A(\boldsymbol{\eta}) - \mathsf{D}_{\boldsymbol{\eta}}\{\mathsf{D}_{\boldsymbol{\eta}}\,A(\boldsymbol{\eta})\,\boldsymbol{\eta}\} = \mathsf{D}_{\boldsymbol{\eta}}A(\boldsymbol{\eta}) - \mathsf{D}_{\boldsymbol{\eta}}\{\boldsymbol{\eta}^T\,\mathsf{D}_{\boldsymbol{\eta}}\,A(\boldsymbol{\eta})^T\}.$$

Next,

$$\begin{aligned} d\{\boldsymbol{\eta}^T\,\mathsf{D}_{\boldsymbol{\eta}}A(\boldsymbol{\eta})^T\} &= (d\boldsymbol{\eta})^T\,\mathsf{D}_{\boldsymbol{\eta}}A(\boldsymbol{\eta})^T + \boldsymbol{\eta}^T\, d\{\mathsf{D}_{\boldsymbol{\eta}}A(\boldsymbol{\eta})^T\} \\ &= \mathsf{D}_{\boldsymbol{\eta}}A(\boldsymbol{\eta})d\boldsymbol{\eta} + \boldsymbol{\eta}^T\,\mathsf{D}_{\boldsymbol{\eta}}\,\{\mathsf{D}_{\boldsymbol{\eta}}A(\boldsymbol{\eta})^T\}d\boldsymbol{\eta} \\ &= \{\mathsf{D}_{\boldsymbol{\eta}}A(\boldsymbol{\eta}) + \boldsymbol{\eta}^T\,\mathsf{H}_{\boldsymbol{\eta}}A(\boldsymbol{\eta})\}\, d\boldsymbol{\eta} \end{aligned}$$

and so

$$\mathsf{D}_{\boldsymbol{\eta}}\{\boldsymbol{\eta}^T \mathsf{D}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})^T\} = \mathsf{D}_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta}).$$

Hence,

$$\mathsf{D}_{\boldsymbol{\eta}} \operatorname{Entropy}\{q(\boldsymbol{x};\boldsymbol{\eta})\} = \mathsf{D}_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) - \{\mathsf{D}_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})\} = -\boldsymbol{\eta}^T \mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta}).$$

Lemma 2 *Let s be a differentiable scalar-valued function of $\boldsymbol{x} \in \mathbb{R}^d$ and let $\boldsymbol{u} \in \mathbb{R}^k$ be one-to-one transformation of $\boldsymbol{x}$. Then*

$$\mathsf{D}_{\boldsymbol{x}}\, s = \left(\mathsf{D}_{\boldsymbol{u}}\, s\right)\left(\mathsf{D}_{\boldsymbol{x}}\, \boldsymbol{u}\right).$$

Proof of Lemma 2

Lemma 2 is a restatement of Theorem 8, Chapter 5, of Magnus and Neudecker (1999).

Lemma 3 *Let $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ have a d-dimensional Multivariate Normal distribution. The natural statistic is*

$$\boldsymbol{T}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \\ \operatorname{vech}(\boldsymbol{x}\boldsymbol{x}^T) \end{bmatrix}$$

and corresponding mean parameter is $\boldsymbol{\tau} \equiv E\{\boldsymbol{T}(\boldsymbol{x})\}$. Then

$$\mathsf{D}_{\boldsymbol{\tau}} \begin{bmatrix} \boldsymbol{\mu} \\ \operatorname{vec}(\boldsymbol{\Sigma}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -(\boldsymbol{I} + \boldsymbol{K}_d)(\mu \otimes \boldsymbol{I}) & \boldsymbol{D}_d \end{bmatrix}.$$

Proof of Lemma 3

The transformation from the common parameters to the mean parameters is

$$\boldsymbol{\tau} \equiv \begin{bmatrix} \boldsymbol{\tau}_1 \\ \boldsymbol{\tau}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \operatorname{vech}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T) \end{bmatrix}$$

and the inverse transformation is easily shown to be

$$\begin{bmatrix} \boldsymbol{\mu} \\ \operatorname{vec}(\boldsymbol{\Sigma}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\tau}_1 \\ \boldsymbol{D}_d\boldsymbol{\tau}_2 - \operatorname{vec}(\boldsymbol{\tau}_1\boldsymbol{\tau}_1^T) \end{bmatrix}.$$

Hence

$$\mathsf{D}_{\boldsymbol{\tau}} \begin{bmatrix} \boldsymbol{\mu} \\ \operatorname{vec}(\boldsymbol{\Sigma}) \end{bmatrix} = \begin{bmatrix} \mathsf{D}_{\boldsymbol{\tau}_1}\boldsymbol{\mu} & \mathsf{D}_{\boldsymbol{\tau}_2}\boldsymbol{\mu} \\ \mathsf{D}_{\boldsymbol{\tau}_1}\operatorname{vec}(\boldsymbol{\Sigma}) & \mathsf{D}_{\boldsymbol{\tau}_2}\operatorname{vec}(\boldsymbol{\Sigma}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\mathsf{D}_{\boldsymbol{\tau}_1}\operatorname{vec}(\boldsymbol{\tau}_1\boldsymbol{\tau}_1^T) & \boldsymbol{D}_d \end{bmatrix}.$$

To obtain an explicit expression for the bottom left-hand block we note that

$$d\operatorname{vec}(\boldsymbol{\tau}_1\boldsymbol{\tau}_1^T) = (\boldsymbol{I} + \boldsymbol{K}_d)\operatorname{vec}\big\{(d\boldsymbol{\tau}_1)\boldsymbol{\tau}_1^T\big\}.$$

Then, with the help of (52),

$$\operatorname{vec}\big\{(d\boldsymbol{\tau}_1)\boldsymbol{\tau}_1^T\big\} = \operatorname{vec}\big\{\boldsymbol{I}(d\boldsymbol{\tau}_1)\boldsymbol{\tau}_1^T\big\} = (\boldsymbol{\tau}_1 \otimes \boldsymbol{I})\, d\boldsymbol{\tau}_1.$$

Hence,

$$\mathsf{D}_{\boldsymbol{\tau}_1}\mathrm{vec}(\boldsymbol{\Sigma}) = (\boldsymbol{I}+\boldsymbol{K}_d)(\boldsymbol{\tau}_1\otimes\boldsymbol{I}) = (\boldsymbol{I}+\boldsymbol{K}_d)(\boldsymbol{\mu}\otimes\boldsymbol{I})$$

and the lemma follows immediately.

Lemma 4 *Let*

$$\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

denote the d-variate $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ density function. Then

$$\mathrm{vec}^{-1}\left\{\mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma})}\,\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})^T\right\} = \tfrac{1}{2}\mathsf{H}_{\boldsymbol{\mu}}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}).$$

Proof of Lemma 4

First note that

$$(2\pi)^{d/2}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-1/2}\exp[-\tfrac{1}{2}\mathrm{tr}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\}].$$

Then, using the identity $\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{B}) = \mathrm{vec}(\boldsymbol{A})^T\mathrm{vec}(\boldsymbol{B})$,

$$\begin{aligned}
(2\pi)^{d/2}d_{\boldsymbol{\Sigma}}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= (d_{\boldsymbol{\Sigma}}|\boldsymbol{\Sigma}|^{-1/2})\exp[-\tfrac{1}{2}\mathrm{tr}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\}] \\
&\quad +|\boldsymbol{\Sigma}|^{-1/2}\Big(d_{\boldsymbol{\Sigma}}\exp[-\tfrac{1}{2}\mathrm{tr}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\}]\Big) \\
&= -\tfrac{1}{2}|\boldsymbol{\Sigma}|^{-3/2}|\boldsymbol{\Sigma}|\mathrm{tr}(\boldsymbol{\Sigma}^{-1}d_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma})\exp[-\tfrac{1}{2}\mathrm{tr}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\}] \\
&\quad +|\boldsymbol{\Sigma}|^{-1/2}\exp[-\tfrac{1}{2}\mathrm{tr}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\}] \\
&\quad\quad \times[-\tfrac{1}{2}\mathrm{tr}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T d_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\}] \\
&= -\tfrac{1}{2}(2\pi)^{d/2}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Sigma}^{-1})^T\,d\mathrm{vec}(\boldsymbol{\Sigma}) \\
&\quad -\tfrac{1}{2}(2\pi)^{d/2}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\mathrm{tr}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\} \\
&= -\tfrac{1}{2}(2\pi)^{d/2}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Sigma}^{-1})^T\,d\mathrm{vec}(\boldsymbol{\Sigma}) \\
&\quad +\tfrac{1}{2}(2\pi)^{d/2}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\,\mathrm{vec}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\}^T d\mathrm{vec}(\boldsymbol{\Sigma}).
\end{aligned}$$

Therefore, by Theorem 6, Chapter 5, of Magnus and Neudecker (1999),

$$\mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma})}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \tfrac{1}{2}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\mathrm{vec}[\boldsymbol{\Sigma}^{-1}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}-\boldsymbol{I}\}]^T.$$

Also,

$$\begin{aligned}
|\boldsymbol{\Sigma}|^{1/2}(2\pi)^{d/2}d_{\boldsymbol{\mu}}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \exp[-\tfrac{1}{2}\mathrm{tr}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\}] \\
&\quad\quad \times[-\tfrac{1}{2}\mathrm{tr}\{d_{\boldsymbol{\mu}}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\}\boldsymbol{\Sigma}^{-1}\}] \\
&= |\boldsymbol{\Sigma}|^{1/2}(2\pi)^{d/2}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\mathrm{tr}\{(\boldsymbol{x}-\boldsymbol{\mu})(d\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\} \\
&= |\boldsymbol{\Sigma}|^{1/2}(2\pi)^{d/2}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T\,d\boldsymbol{\mu}
\end{aligned}$$

which simplifies to

$$d_{\boldsymbol{\mu}}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T\, d\boldsymbol{\mu}.$$

The second differential with respect to $\boldsymbol{\mu}$ is then

$$\begin{aligned} d^2_{\boldsymbol{\mu}}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \{d_{\boldsymbol{\mu}}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T\, d\boldsymbol{\mu} + \phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(-d\boldsymbol{\mu})^T\, d\boldsymbol{\mu} \\ &= \{\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T\, d\boldsymbol{\mu}\}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T\, d\boldsymbol{\mu} + \phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(-d\boldsymbol{\mu})^T\, d\boldsymbol{\mu} \\ &= (d\boldsymbol{\mu})^T\Big(\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})[\boldsymbol{\Sigma}^{-1}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}\}]\Big)\, d\boldsymbol{\mu}. \end{aligned}$$

Hence, using Theorem 6, Chapter 6, of Magnus and Neudecker (1999)

$$\mathsf{H}_{\boldsymbol{\mu}}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})[\boldsymbol{\Sigma}^{-1}\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}\}] = 2\mathrm{vec}^{-1}\Big\{\mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma})}\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})^T\Big\}.$$

A.6 Derivation of Result 2

To make the derivation less cumbersome we will suppress the subscripts on the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. As in Wand (2014) we work with the natural statistic and natural parameter pair

$$\boldsymbol{T}(\boldsymbol{\phi}) = \begin{bmatrix} \boldsymbol{\phi} \\ \mathrm{vech}(\boldsymbol{\phi}\boldsymbol{\phi}^T) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{D}_d\mathrm{vec}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix}.$$

The mean parameter vector is

$$\boldsymbol{\tau} = E\{\boldsymbol{T}(\boldsymbol{\phi})\} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathrm{vech}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T) \end{bmatrix}.$$

In Lemma 3 in Appendix A.5 we show that

$$\mathsf{D}_{\boldsymbol{\tau}}\begin{bmatrix} \boldsymbol{\mu} \\ \mathrm{vec}(\boldsymbol{\Sigma}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -(\boldsymbol{I} + \boldsymbol{K}_d)(\boldsymbol{\mu}\otimes\boldsymbol{I}) & \boldsymbol{D}_d \end{bmatrix}$$

and so Result 1(d) becomes

$$\begin{aligned} &\begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\,\boldsymbol{D}_d^T\,\mathrm{vec}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix} \\ &\qquad = \begin{bmatrix} \boldsymbol{I} & -(\boldsymbol{\mu}^T\otimes\boldsymbol{I})(\boldsymbol{I}+\boldsymbol{K}_d) \\ \boldsymbol{0} & \boldsymbol{D}_d^T \end{bmatrix}\begin{bmatrix} \mathsf{D}_{\boldsymbol{\mu}}\mathrm{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T \\ \mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma})}\mathrm{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T \end{bmatrix} \end{aligned} \tag{55}$$

where we have used the fact that $\boldsymbol{K}_d^T = \boldsymbol{K}_d$. Using (52), and the fact that

$$\mathrm{vec}^{-1}\big[\mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma})}\mathrm{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T\big]$$

is symmetric, the first component of (55) is equivalent to

$$\begin{aligned}
\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \mathsf{D}_{\boldsymbol{\mu}}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T \\
&\qquad -(\boldsymbol{\mu}^T\otimes\boldsymbol{I})(\boldsymbol{I}+\boldsymbol{K}_d)\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T \\
&= \mathsf{D}_{\boldsymbol{\mu}}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T \\
&\qquad -(\boldsymbol{\mu}^T\otimes\boldsymbol{I})(\boldsymbol{I}+\boldsymbol{K}_d)\text{vec}\left(\text{vec}^{-1}\left[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T\right]\right) \\
&= \mathsf{D}_{\boldsymbol{\mu}}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T \\
&\qquad -2(\boldsymbol{\mu}^T\otimes\boldsymbol{I})\text{vec}\left(\text{vec}^{-1}\left[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T\right]\right) \\
&= \mathsf{D}_{\boldsymbol{\mu}}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T \\
&\qquad -2\,\text{vec}\left(\text{vec}^{-1}\left[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T\right]\boldsymbol{\mu}\right) \\
&= \mathsf{D}_{\boldsymbol{\mu}}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T \\
&\qquad -2\,\text{vec}^{-1}\left[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T\right]\boldsymbol{\mu}.
\end{aligned} \tag{56}$$

The second component (55) is equivalent to

$$-\tfrac{1}{2}\,\boldsymbol{D}_d^T\,\text{vec}(\boldsymbol{\Sigma}) = \boldsymbol{D}_d^T\left[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}\right]^T$$

which, under the constraint that $\boldsymbol{\Sigma}$ is symmetric, is equivalent to

$$\boldsymbol{\Sigma}^{-1} = -2\,\text{vec}^{-1}\left[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T\right]. \tag{57}$$

In view of relationships (56) and (57), the natural fixed-point iteration scheme becomes

$$\begin{cases}
\boldsymbol{\Sigma}_{\text{new}}^{-1}\boldsymbol{\mu}_{\text{new}} &= \left[\mathsf{D}_{\boldsymbol{\mu}}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}\right]^T_{\boldsymbol{\mu}=\boldsymbol{\mu}_{\text{old}},\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{\text{old}}} \\
&\qquad -2\,\text{vec}^{-1}\left(\left[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}\right]^T_{\boldsymbol{\mu}=\boldsymbol{\mu}_{\text{old}},\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{\text{old}}}\right)\boldsymbol{\mu}_{\text{old}} \\
\boldsymbol{\Sigma}_{\text{new}} &= \left\{-2\,\text{vec}^{-1}\left(\left[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}\right]^T_{\boldsymbol{\mu}=\boldsymbol{\mu}_{\text{old}},\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{\text{old}}}\right)\right\}^{-1}
\end{cases}$$

where $(\boldsymbol{\mu}_{\text{old}},\boldsymbol{\Sigma}_{\text{old}})$ and $(\boldsymbol{\mu}_{\text{new}},\boldsymbol{\Sigma}_{\text{new}})$, respectively, denote the old and new values of $(\boldsymbol{\mu},\boldsymbol{\Sigma})$. The following simplification ensues:

$$\begin{cases}
\boldsymbol{\mu}_{\text{new}} &= \boldsymbol{\mu}_{\text{old}} + \boldsymbol{\Sigma}_{\text{new}}\left[\mathsf{D}_{\boldsymbol{\mu}}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}\right]^T_{\boldsymbol{\mu}=\boldsymbol{\mu}_{\text{old}},\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{\text{old}}} \\
\boldsymbol{\Sigma}_{\text{new}} &= \left\{-2\,\text{vec}^{-1}\left(\left[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}\right]^T_{\boldsymbol{\mu}=\boldsymbol{\mu}_{\text{old}},\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{\text{old}}}\right)\right\}^{-1}
\end{cases}$$

which is equivalent to the following updating scheme:

$$\begin{cases}
\boldsymbol{v} &\leftarrow \mathsf{D}_{\boldsymbol{\mu}}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T \\
\boldsymbol{\Sigma} &\leftarrow \left(-2\,\text{vec}^{-1}\left[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T\right]\right)^{-1} \\
\boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} + \boldsymbol{\Sigma}\,\boldsymbol{v}
\end{cases} \tag{58}$$

as given in Wand (2014). However, from Lemma 4 in Appendix A.5 (see also Appendix A of Opper and Archambeau, 2009) and the fact that NonEntropy$\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}$ is an expectation with respect to the $N(\mu,\Sigma)$ density function we have

$$\text{vec}^{-1}\big[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}^T\big] = \tfrac{1}{2}\mathsf{H}_{\boldsymbol{\mu}}\text{NonEntropy}\{q(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma})\}$$

which leads to a somewhat more elegant alternative to (58) given in Result 2.

A.7 Derivation of the Derivative Matrix of $\boldsymbol{g}_{\text{Ex2}}$

From Section 4.3, the fixed-point iteration updating function is of the form

$$\boldsymbol{g}_{\text{Ex2}}\left(\begin{bmatrix}\boldsymbol{\mu}\\ \text{vech}(\boldsymbol{\Sigma})\end{bmatrix}\right) \equiv \begin{bmatrix}\boldsymbol{g}_{\text{Ex2},\boldsymbol{\mu}}\\ \text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})\end{bmatrix}$$

where

$$\begin{aligned}\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}} &\equiv \big[\boldsymbol{C}^T\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{C}+\boldsymbol{M}\big]^{-1},\\ \boldsymbol{g}_{\text{Ex2},\boldsymbol{\mu}} &\equiv \boldsymbol{\mu}+\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}[\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}-\boldsymbol{M}\boldsymbol{\mu}]\\ \text{and}\quad \boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma}) &\equiv \exp\{\boldsymbol{C}\boldsymbol{\mu}+\tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T)\}.\end{aligned}$$

Note that the dependence of $\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}$ and $\boldsymbol{g}_{\text{Ex2}}$ on $\boldsymbol{y}$, $\boldsymbol{C}$ and $\boldsymbol{M}$ is suppressed here.

The derivative matrix of $\boldsymbol{g}_{\text{Ex2}}$ with respect to $[\boldsymbol{\mu}\ \text{vech}(\boldsymbol{\Sigma})]^T$ is

$$\mathsf{D}_{\left[\begin{smallmatrix}\boldsymbol{\mu}\\ \text{vech}(\boldsymbol{\Sigma})\end{smallmatrix}\right]}\boldsymbol{g}_{\text{Ex2}} = \begin{bmatrix}\mathsf{D}_{\boldsymbol{\mu}}\,\boldsymbol{g}_{\text{Ex2},\boldsymbol{\mu}} & \mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\boldsymbol{g}_{\text{Ex2},\boldsymbol{\mu}}\\ \mathsf{D}_{\boldsymbol{\mu}}\,\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}) & \mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})\end{bmatrix}.$$

We now give explicit expressions for each of these four components of the derivative matrix. It is more efficient, notationally, to first obtain expressions for the derivatives of $\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})$.

A.7.1 Expression for $\mathsf{D}_{\boldsymbol{\mu}}\,\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})$

$$\mathsf{D}_{\boldsymbol{\mu}}\,\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}) = -\boldsymbol{D}_{p+K}^{+}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\otimes\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})\boldsymbol{Q}(\boldsymbol{C})^T\,\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{C} \qquad (59)$$

Derivation:

Using the second rule in Section 3.3.5 of Wand (2002), (52) and (51),

$$\begin{aligned}d_{\boldsymbol{\mu}}\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}) &= \boldsymbol{D}_{p+K}^{+}\,d_{\boldsymbol{\mu}}\text{vec}\Big([\boldsymbol{C}^T\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\boldsymbol{C}\}+\boldsymbol{M}]^{-1}\Big)\\ &= -\boldsymbol{D}_{p+K}^{+}\text{vec}\Big\{\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\Big(d_{\boldsymbol{\mu}}[\boldsymbol{C}^T\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{C}+\boldsymbol{M}]\Big)\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\Big\}\\ &= -\boldsymbol{D}_{p+K}^{+}\Big(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\otimes\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\Big)\,\text{vec}[\boldsymbol{C}^T\text{diag}\{d_{\boldsymbol{\mu}}\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{C}].\end{aligned}$$

From Theorem 2(b) of Wand (2014),

$$\text{vec}[\boldsymbol{C}^T\text{diag}\{d_{\boldsymbol{\mu}}\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{C}] = \boldsymbol{Q}(\boldsymbol{C})^T d_{\boldsymbol{\mu}}\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma}).$$

Lastly, we use the chain rule in Section 3.3.2 of Wand (2002) to get

$$d_{\boldsymbol{\mu}}\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = d_{\boldsymbol{\mu}}\exp\{\boldsymbol{C}\boldsymbol{\mu} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T)\} = \text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{C}\,d\mu.$$

Combining, we then obtain

$$d_{\boldsymbol{\mu}}\text{vec}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}) = -\boldsymbol{D}^{+}_{p+K}\Big(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\otimes\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\Big)\,\boldsymbol{Q}(\boldsymbol{C})^T\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{C}\,d\mu$$

and the stated expression then follow from Theorem 6, Chapter 5, of Magnus and Neudecker (1999).

A.7.2 Expression for $\mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})$

$$\begin{aligned}\mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}) \;&=\; -\tfrac{1}{2}\boldsymbol{D}^{+}_{p+K}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\otimes\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})\\ &\qquad\times\boldsymbol{Q}(\boldsymbol{C})^T\,\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{Q}(\boldsymbol{C})\boldsymbol{D}_{p+K}.\end{aligned} \tag{60}$$

Derivation:

The derivation is similar to that for $\mathsf{D}_{\boldsymbol{\mu}}\,\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}$. It differs in that it requires $d_{\boldsymbol{\Sigma}}\,\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ rather than $d_{\boldsymbol{\mu}}\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. This entails

$$d_{\boldsymbol{\Sigma}}\,\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = d_{\boldsymbol{\Sigma}}\exp\{\boldsymbol{C}\boldsymbol{\mu} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T)\} = \tfrac{1}{2}\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}d_{\boldsymbol{\Sigma}}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T).$$

But Theorem 2(a) of Wand (2014) gives

$$d_{\boldsymbol{\Sigma}}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T) = \boldsymbol{Q}(\boldsymbol{C})d_{\boldsymbol{\Sigma}}\text{vec}(\boldsymbol{\Sigma}) = \boldsymbol{Q}(\boldsymbol{C})\boldsymbol{D}_{p+K}\,d_{\boldsymbol{\Sigma}}\text{vech}(\boldsymbol{\Sigma})$$

which leads to the stated result.

A.7.3 Expression for $\mathsf{D}_{\boldsymbol{\mu}}\,\boldsymbol{g}_{\text{Ex2},\boldsymbol{\mu}}$

$$\mathsf{D}_{\boldsymbol{\mu}}\,\boldsymbol{g}_{\text{Ex2},\boldsymbol{\mu}} = \big([\,\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\} - \boldsymbol{M}\boldsymbol{\mu}]^T\otimes\boldsymbol{I}\big)\boldsymbol{D}_{p+K}\mathsf{D}_{\boldsymbol{\mu}}\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})$$

where $\mathsf{D}_{\boldsymbol{\mu}}\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})$ is given by (59).

Derivation:

Using the second rule in Section 3.3.4 of Wand (2002) for differentiation of matrix products,

$$\begin{aligned}d_{\boldsymbol{\mu}}\boldsymbol{g}_{\text{Ex2},\boldsymbol{\mu}} \;&=\; d\boldsymbol{\mu} + d_{\boldsymbol{\mu}}\Big(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}[\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\} - \boldsymbol{M}\boldsymbol{\mu}]\Big)\\ &=\; d\boldsymbol{\mu} + \big(d_{\boldsymbol{\mu}}\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\big)[\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\} - \boldsymbol{M}\boldsymbol{\mu}]\\ &\qquad -\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}[\boldsymbol{C}^T d_{\boldsymbol{\mu}}\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma}) + \boldsymbol{M}\,d\boldsymbol{\mu}]\\ &=\; d\boldsymbol{\mu} + \text{vec}\big\{\boldsymbol{I}\big(d_{\boldsymbol{\mu}}\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\big)[\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\} - \boldsymbol{M}\boldsymbol{\mu}]\big\}\\ &\qquad -\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}[\boldsymbol{C}^T\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{C} + \boldsymbol{M}]\,d\boldsymbol{\mu},\end{aligned}$$

where we have used the fact that $\big(d_{\boldsymbol{\mu}}\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\big)[\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}-\boldsymbol{M}\boldsymbol{\mu}]$ is a column vector. Application of (52) leads to

$$\begin{aligned}
d_{\boldsymbol{\mu}}\boldsymbol{g}_{\text{Ex2},\boldsymbol{\mu}} &= d\boldsymbol{\mu} + \big([\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}-\boldsymbol{M}\boldsymbol{\mu}]^T\otimes\boldsymbol{I}\big)d_{\boldsymbol{\mu}}\text{vec}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}) \\
&\quad -\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}[\boldsymbol{C}^T\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{C}+\boldsymbol{M}]\,d\boldsymbol{\mu} \\
&= \Big\{\boldsymbol{I} + \big([\,\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}-\boldsymbol{M}\boldsymbol{\mu}]^T\otimes\boldsymbol{I}\big)\mathsf{D}_{\boldsymbol{\mu}}\text{vec}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}) \\
&\quad -\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\boldsymbol{G}^{-1}_{\text{Ex2},\boldsymbol{\Sigma}}\Big\}\,d\boldsymbol{\mu} \\
&= \big([\,\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}-\boldsymbol{M}\boldsymbol{\mu}]^T\otimes\boldsymbol{I}\big)\boldsymbol{D}_{p+K}\mathsf{D}_{\boldsymbol{\mu}}\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})\,d\boldsymbol{\mu}.
\end{aligned}$$

The given expression follows from Theorem 6, Chapter 5, of Magnus and Neudecker (1999).

A.7.4 Expression for $\mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\boldsymbol{g}_{\text{Ex2},\boldsymbol{\mu}}$

$$\begin{aligned}
\mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\boldsymbol{g}_{\text{Ex2},\boldsymbol{\mu}} &= \Big([\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}-\boldsymbol{M}\boldsymbol{\mu}]^T\otimes\boldsymbol{I}\Big)\boldsymbol{D}_{p+K}\mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}) \\
&\quad -\tfrac{1}{2}\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\boldsymbol{C}^T\,\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{Q}(\boldsymbol{C})\boldsymbol{D}_{p+K}
\end{aligned}$$

where $\mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})$ is given by (60).

Derivation:

Dealing with matrix products via the second rule in Section 3.3.4 of Wand (2002) we obtain

$$\begin{aligned}
d_{\boldsymbol{\Sigma}}\boldsymbol{g}_{\text{Ex2},\boldsymbol{\mu}} &= d_{\boldsymbol{\Sigma}}\Big(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}[\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}-\boldsymbol{M}\boldsymbol{\mu}]\Big) \\
&= (d_{\boldsymbol{\Sigma}}\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})[\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}-\boldsymbol{M}\boldsymbol{\mu}] \\
&\quad -\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\boldsymbol{C}^T\,d_{\boldsymbol{\Sigma}}\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&= \text{vec}\Big(\boldsymbol{I}(d_{\boldsymbol{\Sigma}}\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})[\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}-\boldsymbol{M}\boldsymbol{\mu}]\Big) \\
&\quad -\tfrac{1}{2}\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\boldsymbol{C}^T\,\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{Q}(\boldsymbol{C})\,d\text{vec}(\boldsymbol{\Sigma}) \\
&= \big([\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}-\boldsymbol{M}\boldsymbol{\mu}]^T\otimes\boldsymbol{I}\big)\,\boldsymbol{D}_{p+K}d\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}) \\
&\quad -\tfrac{1}{2}\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\boldsymbol{C}^T\,\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{Q}(\boldsymbol{C})\,\boldsymbol{D}_{p+K}d\text{vech}(\boldsymbol{\Sigma}) \\
&= \big([\boldsymbol{C}^T\{\boldsymbol{y}-\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}-\boldsymbol{M}\boldsymbol{\mu}]^T\otimes\boldsymbol{I}\big) \\
&\quad \times\boldsymbol{D}_{p+K}\mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\,\text{vech}(\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}})d\text{vech}(\boldsymbol{\Sigma}) \\
&\quad -\tfrac{1}{2}\boldsymbol{G}_{\text{Ex2},\boldsymbol{\Sigma}}\boldsymbol{C}^T\,\text{diag}\{\boldsymbol{\omega}(\boldsymbol{\mu},\boldsymbol{\Sigma})\}\boldsymbol{Q}(\boldsymbol{C})\,\boldsymbol{D}_{p+K}d\text{vech}(\boldsymbol{\Sigma}).
\end{aligned}$$

Once again we call upon Theorem 6, Chapter 5, of Magnus and Neudecker (1999) to complete the derivation.

References

A.S. Ackleh, E.J. Allen, R.B. Hearfott, and P. Seshaiyer. *Classical and Modern Numerical Analysis: Theory, Methods and Practice*. Chapman & Hall, CRC Press, Boca Raton, Florida, 2010.

S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor. Gaussian process approximations of stochastic differential equations. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 1:1–16, 2007.

A. Azzalini and A. Dalla Valle. The Multivariate Skew-Normal distribution. *Biometrika*, 83:715–726, 1996.

D. Barber and C.M. Bishop. Ensemble learning for multi-layer networks. In M.I. Jordan, K.J. Kearns, and S.A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 395–401, 1997.

M.J. Beal and Z. Ghahramani. Variational Bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 4:793–832, 2006.

C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

L. Bottou. Stochastic learning. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures on Machine Learning*, pages 146–168, 2004.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

E. Challis and D. Barber. Gaussian Kullback-Leibler approximate inference. *Journal of Machine Learning Research*, 14:2239–2286, 2013.

Y.H. Dai and Y. Yuan. A nonlinear conjugate gradient method with a strong global convergence property. *SIAM Journal on Optimization*, 10:177–182, 1999.

S.J. Gershman, M.D. Hoffman, and D.M. Blei. Nonparametric variational inference. In *Proceedings of the Twenty Ninth International Conference on Machine Learning*, pages 633–670, 2012.

G.H. Givens and J.A. Hoeting. *Computational Statistics*. Wiley, Hoboken, New Jersey, 2005.

M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49:409–436, 1952.

G.E. Hinton and D. van Camp. Keeping neural networks simple by minimizing description length of the weights. In L. Pitt, editor, *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pages 5–13, 1993.

M.D. Hoffman, D.M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

A. Honkela and H. Valpola. Unsupervised variational Bayesian learning of nonlinear models. *Advances in Neural Information Processing Systems*, 17:593–600, 2005.

A. Honkela, H. Valpola, A. Ilin, and J. Karhunen. Blind separation of nonlinear mixtures by variational Bayesian learning. *Digital Signal Processing*, 17:914–934, 2007.

A. Honkela, M. Tornio, T. Raiko, and J. Karhunen. Natural conjugate gradient in variational inference. In M. Ishikawa, K. Doya, H. Miyamoto, and T. Yamakawa, editors, *Proceedings of the Fourteenth International Conference on Neural Information Processing*, pages 305–314, 2008.

A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *Journal of Machine Learning Research*, 11:3235–3268, 2010.

D.A. Knowles and T.P. Minka. Non-conjugate message passing for multinomial and binary regression. In J. Shawe-Taylor, R.S. Zamel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, 2011.

H. Lappalainen and A. Honkela. Bayesian nonlinear independent component analysis by multilayer perceptrons. In M. Girolami, editor, *Advances in Independent Component Analysis*, pages 93–121, 2000.

D.G. Luenberger and Y. Ye. *Linear and Nonlinear Programming, Third Edition*. Springer, New York, 2008.

J. Luts and M.P. Wand. Variational inference for count response semiparametric regression. *Bayesian Analysis*, 10:991–1023, 2015.

J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition*. Wiley, Chichester, UK, 1999.

M. Menictas and M.P. Wand. Variational inference for heteroscedastic semiparametric regression. *Australian and New Zealand Journal of Statistics*, 57:119–138, 2015.

M.K. Murray and J.W. Rice. *Differential Geometry and Statistics*. Chapman & Hall, London, 1993.

J.A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

S.E. Neville, J.T. Ormerod, and M.P. Wand. Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electronic Journal of Statistics*, 8:1113–1151, 2014.

H. Nickisch and C.E. Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, 2008.

M. Opper and C. Archambeau. The variational Gaussian approximation revisited. *Neural Computation*, 21:786–792, 2009.

J.T. Ormerod and M.P. Wand. Explaining variational approximations. *The American Statistician*, 64:140–153, 2010.

J.M. Ortega. *Numerical Analysis: A Second Course*. SIAM, Philadelphia, 1990.

T. Pham, J.T. Ormerod, and M.P. Wand. Mean field variational Bayesian inference for nonparametric regression with measurement error. *Computational Statistics and Data Analysis*, 68:375–387, 2013.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flanney. *Numerical Recipes: The Art of Scientific Computing, Third Edition*. Cambridge University Press, New York, 2007.

T. Raiko, H. Valpola, M. Harva, and J. Karhunen. Building blocks for variational Bayesian learning of latent variable models. *Journal of Machine Learning Research*, 8:155–201, 2007.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `http://www.R-project.org/`.

T. Salimans and D.A. Knowles. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8:837–882, 2013.

M. Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13: 1649–1681, 2001.

L.S.L. Tan and D.J. Nott. Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science*, 28:168–188, 2013.

M.J. Wainwright and M.I. Jordan. Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

M.P. Wand. Vector differential calculus in statistics. *The American Statistician*, 56:55–62, 2002.

M.P. Wand. Fully simplified Multivariate Normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research*, 15:1351–1369, 2014.

M.P. Wand, J.T. Ormerod, S.A. Padoan, and R. Frühwirth. Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, 6:847–900, 2011.

B. Wang and D.M. Titterington. Inadequacy of interval estimates corresponding to variational Bayes approximations. In Z. Ghahramani and R. Cowell, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 373–380, 2005.

O. Zobay. Variational Bayesian inference with Gaussian-mixture approximations. *Electronic Journal of Statistics*, 8:355–389, 2014.