

Moment Propagation

John T. Ormerod¹, Weichang Yu², Mohammad Javad Davoudabadi^{*1},
and Yuhao Li¹

¹*School of Mathematics and Statistics, University of Sydney, Australia;*

²*School of Mathematics and Statistics, University of Melbourne, Parkville, Australia.*

Abstract

Mean-field variational Bayes is a fast and scalable approach to approximate Bayesian inference, but its independence assumptions often lead to underestimated posterior uncertainty. We introduce moment propagation (MP), a framework for improving marginal posterior approximations by propagating conditional posterior moment information between parameter blocks and matching these moments within convenient approximating families. In conjugate settings, MP identifies variance terms omitted by mean-field approximations and uses them to construct corrected marginal updates. We develop both mean-field and Gaussian variants of MP, the latter using conditional Gaussian structure to obtain low-dimensional local updates for non-conjugate models. We establish consistency and asymptotic variance correctness for two-component models, and show that MP recovers exact marginal posteriors for linear regression and multivariate normal models. We also derive algorithms for multivariate normal models with missing data, probit regression, generalized linear models, and Bayesian Lasso regression. Numerical experiments show that MP, particularly Gaussian MP, can substantially improve marginal posterior accuracy over standard variational approximations while remaining much faster than MCMC.

Keywords: Approximate Bayesian Inference, Variational Bayes, Moment Propagation.

^{*}mohammad.davoudabadi@sydney.edu.au

1 Introduction

Bayesian inference provides a coherent framework for uncertainty quantification, but exact posterior inference is available only in relatively simple models. Markov chain Monte Carlo (MCMC) methods remain a default benchmark for posterior approximation because, under suitable conditions, they can generate asymptotically exact samples from the posterior distribution. However, MCMC can be computationally expensive for large datasets, high-dimensional or complex models, and repeated analyses. This has motivated approximate Bayesian inference (ABI) methods that trade some accuracy for substantial computational gains.

Variational Bayes (VB), also known as variational inference (VI), is one of the most widely used families of ABI methods. VB converts posterior approximation into an optimization problem by selecting a tractable class of approximating densities and finding the member of this class that is closest to the posterior distribution in Kullback–Leibler divergence (Kullback and Leibler, 1951); see, for example, Bishop (2006), Ormerod and Wand (2010), Murphy (2013), and Blei et al. (2017). Among VB methods, mean-field variational Bayes (MFVB) is especially popular because it leads to simple coordinate ascent algorithms and can scale well to large problems. MFVB assumes that the posterior approximation factorizes across a partition of the model parameters, often yielding analytically tractable updates in conjugate models.

Despite these advantages, MFVB is well known to produce poorly calibrated uncertainty estimates. Empirical and theoretical studies have shown that MFVB often underestimates posterior variances and ignores important posterior dependence between parameters; see, for example, Wang and Titterton (2005) and Turner et al. (2011). This can lead to overly narrow credible intervals, overconfident posterior rankings, and misleading downstream decisions. As a result, MFVB is often more reliable for fast model exploration and prediction than for inferential tasks requiring accurate marginal posterior uncertainty.

Several approaches have been proposed to improve deterministic Bayesian approximations. Fixed-form variational methods, including Gaussian variational Bayes (GVB), allow richer approximating families than standard mean-field assumptions (Opper and Archambeau, 2009; Challis and Barber, 2013). Expectation propagation (EP) uses local approximations and mo-

ment matching to construct accurate global approximations for many models (Minka, 2001). Laplace and integrated nested Laplace approximations use local curvature and low-dimensional numerical integration to approximate marginal posterior distributions (Rue et al., 2009). Linear response variational Bayes (LRVB) and related covariance correction methods improve uncertainty quantification by studying the sensitivity of MFVB solutions to perturbations (Giordano et al., 2015). These methods have different strengths, but some are model-specific, require expensive optimization, are sensitive to implementation details, or provide corrected moments without directly constructing improved marginal approximating densities.

In this paper we introduce moment propagation (MP), a framework for improving marginal posterior approximations by propagating conditional posterior moment information between parameter blocks. The starting point is the observation that MFVB updates often behave like plug-in approximations: they replace uncertain quantities in a parameter block’s full conditional distribution by their variational expectations. This is computationally convenient, but it omits terms arising from uncertainty in the conditioning variables. In conjugate block models, these missing terms can be identified directly using the laws of total expectation and total covariance. MP uses this insight to construct improved marginal approximations by matching moments obtained from local conditional posterior calculations.

The basic idea is simple. Suppose the parameter vector is partitioned into components. Rather than approximating the conditional distribution of one component given fixed expectations of the others, MP approximates the marginal posterior moments of a component by averaging conditional posterior moments over an approximation to the distribution of the relevant neighbouring components. The resulting moments are then matched to a convenient approximating family. In this way, MP retains much of the blockwise structure of MFVB while explicitly propagating information about posterior uncertainty between parameter blocks.

We develop two main versions of the method. Mean-field moment propagation (MFMP) is closest in spirit to MFVB: it approximates neighbouring components by a product of marginal approximations, but uses conditional posterior moments. This gives a direct correction to MFVB-style variance underestimation in conjugate settings. Gaussian moment propagation (GMP) starts from a joint Gaussian approximation to the posterior and uses its conditional

structure to form low-dimensional local approximations to marginal posterior distributions. These local approximations are then used to update the Gaussian approximation by moment matching, providing a practical route for applying MP in non-conjugate models.

The contributions of this paper are as follows. First, we identify how MFVB can underestimate posterior variances in conjugate block models and use this analysis to motivate moment-based corrections. Second, we introduce MP as a general framework for constructing improved marginal posterior approximations from conditional posterior moments. Third, we develop MFMP and GMP algorithms, showing how they can be implemented using exact conditional moments, low-dimensional numerical integration, and delta-method approximations. Fourth, we establish theoretical properties of MP in two-component conjugate settings, including consistency and asymptotic correctness of posterior variances. Finally, we derive MP algorithms for several archetypal models, including linear regression, multivariate normal models, multivariate normal models with missing data, generalized linear models, and Bayesian Lasso regression.

The examples illustrate both conjugate and non-conjugate uses of MP. Linear regression and multivariate normal models provide exact-recovery benchmarks. Multivariate normal models with missing data demonstrate propagation of latent-variable uncertainty. Probit regression illustrates an auxiliary-variable setting in which not all approximating densities need to be specified explicitly. Generalized linear models and Bayesian Lasso regression demonstrate GMP in non-conjugate settings, including a model with a non-smooth prior.

Our numerical results compare MP-based methods with standard ABI methods. Across the examples considered, MP-based approximations often provide substantially more accurate marginal posterior approximations than MFVB, particularly for posterior standard deviations and marginal posterior densities, while remaining much faster than routine MCMC.

The remainder of the paper is organized as follows. Section 2 reviews VB and identifies the sources of MFVB variance underestimation in conjugate and non-conjugate settings. Section 3 introduces the moment propagation framework, including MFMP, GMP, and hybrid variants. Section 4 derives MP algorithms for the example models. Section 5 presents numerical experiments comparing MP with existing ABI methods. Section 6 concludes. Technical derivations and proofs are provided in the Supplementary Material.

2 Variational Bayes

Let $\mathcal{D}$ denote the observed data and let $\boldsymbol{\theta} \in \mathbb{R}^d$ denote the unknown parameters. Given a likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ and prior density $p(\boldsymbol{\theta})$, the posterior distribution is

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}, \quad p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

For most models of practical interest, the marginal likelihood $p(\mathcal{D})$ is unavailable in closed form and posterior approximation is required.

Variational Bayes (VB), also known as variational inference (VI), approximates $p(\boldsymbol{\theta}|\mathcal{D})$ by choosing a tractable family of densities $\mathcal{Q} = \{q(\boldsymbol{\theta})\}$ and solving $q^\star = \arg \min_{q \in \mathcal{Q}} \text{KL}[q(\boldsymbol{\theta}), p(\boldsymbol{\theta}|\mathcal{D})]$, where $\text{KL}[q(\boldsymbol{\theta}), p(\boldsymbol{\theta}|\mathcal{D})] = \mathbb{E}_q[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathcal{D})]$. For any $q \in \mathcal{Q}$, we have

$$\log p(\mathcal{D}) = \mathbb{E}_q[\log p(\mathcal{D}, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta})] + \text{KL}[q(\boldsymbol{\theta}), p(\boldsymbol{\theta}|\mathcal{D})]. \quad (1)$$

The first term on the right-hand side of (1) is the evidence lower bound (ELBO), $\text{ELBO}_q \equiv \mathbb{E}_q[\log p(\mathcal{D}, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta})]$. Since $\log p(\mathcal{D})$ does not depend on q , minimizing $\text{KL}[q(\boldsymbol{\theta}), p(\boldsymbol{\theta}|\mathcal{D})]$ is equivalent to maximizing ELBO_q .

Two common choices of $\mathcal{Q}$ are especially relevant here. The first is the mean-field family,

$$q(\boldsymbol{\theta}) = \prod_{k=1}^K q_k(\boldsymbol{\theta}_k),$$

where $\{\boldsymbol{\theta}_k\}_{k=1}^K$ is a partition of $\boldsymbol{\theta}$. This leads to mean-field variational Bayes (MFVB). The second is a fixed-form family $q(\boldsymbol{\theta}; \boldsymbol{\eta})$, where the approximating density is specified up to a finite-dimensional variational parameter $\boldsymbol{\eta}$. A common fixed-form choice is a multivariate Gaussian density, often referred to as Gaussian variational Bayes. Intermediate choices, combining structured factorizations and parametric approximations, are also possible (Wang and Blei, 2013; Rohde and Wand, 2016).

The mean-field assumption is computationally attractive because it often leads to simple coordinate ascent updates. However, the factorization removes posterior dependence between

parameter blocks and is a major source of inaccurate posterior uncertainty. The purpose of this section is to make this loss explicit and to identify the moment corrections that motivate the moment propagation framework introduced in Section 3.

Let $M(k) \subseteq \{1, \dots, K\} \setminus \{k\}$ denote the indices of the parameter blocks in the Markov blanket of $\boldsymbol{\theta}_k$. Then $p(\boldsymbol{\theta}_k | \mathcal{D}, \boldsymbol{\theta}_{-k}) = p(\boldsymbol{\theta}_k | \mathcal{D}, \boldsymbol{\theta}_{M(k)})$, where $\boldsymbol{\theta}_{-k}$ denotes all blocks except $\boldsymbol{\theta}_k$. We refer to this conditional distribution as the full conditional distribution of $\boldsymbol{\theta}_k$ and write it as $p(\boldsymbol{\theta}_k | \text{rest}) \equiv p(\boldsymbol{\theta}_k | \mathcal{D}, \boldsymbol{\theta}_{M(k)})$. This conditional distribution plays a central role in both Gibbs sampling and MFVB coordinate ascent updates.

MFVB is particularly convenient when the full conditionals belong to exponential families. Suppose that $p(\boldsymbol{\theta}_k | \text{rest}) = h_k(\boldsymbol{\theta}_k) \exp \left[\mathbf{T}_k(\boldsymbol{\theta}_k)^T \boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)}) - A_k\{\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})\} \right]$, where $\mathbf{T}_k(\boldsymbol{\theta}_k)$ is a vector of sufficient statistics, $\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})$ is the natural parameter, $A_k(\cdot)$ is the log-partition function, and $h_k(\boldsymbol{\theta}_k)$ is the base measure. The natural parameter may depend on $\mathcal{D}$, but we suppress this dependence in the notation. For fixed $\{q_j(\boldsymbol{\theta}_j) : j \neq k\}$, the coordinate ascent variational update is $q_k^*(\boldsymbol{\theta}_k) \propto \exp \left[\mathbb{E}_{q(\boldsymbol{\theta}_{-k})} \{\log p(\mathcal{D}, \boldsymbol{\theta})\} \right]$. In the exponential-family case

$$q_k^*(\boldsymbol{\theta}_k) = h_k(\boldsymbol{\theta}_k) \exp \left[\mathbf{T}_k(\boldsymbol{\theta}_k)^T \hat{\boldsymbol{\eta}}_k - A_k(\hat{\boldsymbol{\eta}}_k) \right], \quad \text{where} \quad \hat{\boldsymbol{\eta}}_k = \mathbb{E}_q[\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})]. \quad (2)$$

Thus, the MFVB update replaces $\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})$ by its expectation under the current variational approximation. Sequentially applying (2) for $k = 1, \dots, K$ yields coordinate ascent variational inference and monotonically increases the ELBO under the usual regularity conditions.

2.1 Variance underestimation in conjugate MFVB updates

The plug-in nature of (2) helps explain why MFVB often underestimates posterior variances. Let $\mathbf{g}_k(\boldsymbol{\eta}_k) = \nabla A_k(\boldsymbol{\eta}_k)$, and $\mathbf{H}_k(\boldsymbol{\eta}_k) = \nabla^2 A_k(\boldsymbol{\eta}_k)$. For an exponential-family distribution with natural parameter $\boldsymbol{\eta}_k$, these quantities are the mean and covariance of the sufficient statistics: $\mathbb{E}[\mathbf{T}_k(\boldsymbol{\theta}_k)] = \mathbf{g}_k(\boldsymbol{\eta}_k)$ and $\mathbb{V}[\mathbf{T}_k(\boldsymbol{\theta}_k)] = \mathbf{H}_k(\boldsymbol{\eta}_k)$. Therefore, under the MFVB update in (2), $\mathbb{E}_q[\mathbf{T}_k(\boldsymbol{\theta}_k)] = \mathbf{g}_k \left[\mathbb{E}_q\{\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})\} \right]$, and

$$\mathbb{V}_q[\mathbf{T}_k(\boldsymbol{\theta}_k)] = \mathbf{H}_k \left[\mathbb{E}_q\{\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})\} \right]. \quad (3)$$

By contrast, the exact posterior moments satisfy the laws of total expectation and total covariance. Conditional on $\boldsymbol{\theta}_{M(k)}$, the full conditional distribution has natural parameter $\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})$. Hence, $\mathbb{E}[\mathbf{T}_k(\boldsymbol{\theta}_k)|\mathcal{D}] = \mathbb{E}_{p(\boldsymbol{\theta}_{M(k)}|\mathcal{D})} [\mathbf{g}_k\{\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})\}]$, and

$$\mathbb{V}[\mathbf{T}_k(\boldsymbol{\theta}_k)|\mathcal{D}] = \mathbb{E}_{p(\boldsymbol{\theta}_{M(k)}|\mathcal{D})} [\mathbf{H}_k\{\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})\}] + \mathbb{V}_{p(\boldsymbol{\theta}_{M(k)}|\mathcal{D})} [\mathbf{g}_k\{\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})\}]. \quad (4)$$

The first term in (4) is the average conditional covariance of the sufficient statistics. The second term is the variation in the conditional mean induced by posterior uncertainty in the neighbouring blocks. Comparing (3) and (4) identifies the source of variance underestimation. If $q(\boldsymbol{\theta}_{M(k)})$ is a good approximation to $p(\boldsymbol{\theta}_{M(k)}|\mathcal{D})$, then the MFVB quantities

$$\mathbf{g}_k [\mathbb{E}_q\{\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})\}] \quad \text{and} \quad \mathbf{H}_k [\mathbb{E}_q\{\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})\}]$$

can be interpreted as plug-in or first-order delta-method approximations to

$$\mathbb{E}_{p(\boldsymbol{\theta}_{M(k)}|\mathcal{D})} [\mathbf{g}_k\{\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})\}] \quad \text{and} \quad \mathbb{E}_{p(\boldsymbol{\theta}_{M(k)}|\mathcal{D})} [\mathbf{H}_k\{\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})\}],$$

respectively. However, the MFVB covariance in (3) has no analogue of the second term in (4), i.e., $\mathbb{V}_{p(\boldsymbol{\theta}_{M(k)}|\mathcal{D})} [\mathbf{g}_k\{\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})\}]$. This missing term is nonnegative definite and represents the uncertainty propagated from the neighbouring parameter blocks into the conditional mean of $\mathbf{T}_k(\boldsymbol{\theta}_k)$. Whenever $\mathbf{T}_k(\boldsymbol{\theta}_k)$ contains $\boldsymbol{\theta}_k$ itself, the same mechanism leads directly to underestimation of the marginal posterior variance of $\boldsymbol{\theta}_k$. This decomposition is the main conjugate motivation for moment propagation. Rather than treating the natural parameter $\boldsymbol{\eta}_k(\boldsymbol{\theta}_{M(k)})$ as fixed at its variational expectation, one can attempt to average conditional moments over an approximation to the posterior distribution of the neighbouring blocks. This restores, at least approximately, the second term in the total covariance decomposition.

2.2 A conditional identity for non-conjugate updates

The preceding argument relies on the full conditional distribution being available in a tractable exponential-family form. For non-conjugate models, the same plug-in interpretation is less

direct, but a related identity suggests a useful route to improvement. Let $q(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k)$ be any conditional density. Applying the decomposition in (1), now conditional on $\boldsymbol{\theta}_k$, gives

$$\log p(\mathcal{D}, \boldsymbol{\theta}_k) = \mathbb{E}_{q(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k)} [\log p(\mathcal{D}, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k)] + \text{KL}[q(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k), p(\boldsymbol{\theta}_{-k}|\mathcal{D}, \boldsymbol{\theta}_k)]. \quad (5)$$

If $q(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k)$ is replaced by a marginal approximation $q(\boldsymbol{\theta}_{-k})$ that does not depend on $\boldsymbol{\theta}_k$, and if the KL term is ignored as a function of $\boldsymbol{\theta}_k$, then (5) leads to the usual MFVB-style update. This approximation discards the posterior dependence of $\boldsymbol{\theta}_{-k}$ on the block currently being updated. Equation (5) therefore suggests a non-conjugate analogue of the conjugate correction above. If $p(\boldsymbol{\theta}_{-k}|\mathcal{D}, \boldsymbol{\theta}_k)$ can be approximated well by a conditional density $q(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k)$, then the first term on the right-hand side of (5) provides an improved local approximation to the marginal log-density of $\boldsymbol{\theta}_k$. In particular, retaining dependence on $\boldsymbol{\theta}_k$ allows information about the remaining parameters to be propagated into the marginal approximation of $\boldsymbol{\theta}_k$.

The calculations above identify two complementary ways that MFVB can lose posterior information. In conjugate models, MFVB omits the contribution of uncertainty in neighbouring blocks to the marginal posterior covariance. In non-conjugate models, MFVB-style updates ignore the conditional dependence of the remaining parameters on the current updated block. Moment propagation uses these observations to construct improved marginal posterior approximations by propagating conditional posterior moment information between parameter blocks.

3 Moment propagation

We now introduce moment propagation (MP), a framework for constructing improved marginal posterior approximations by propagating conditional posterior moment information between parameter blocks. Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$, $\boldsymbol{\theta}_k \in \mathbb{R}^{d_k}$, denote a partition of the parameter vector. Our goal is to construct approximations $q_k(\boldsymbol{\theta}_k) \approx p(\boldsymbol{\theta}_k|\mathcal{D})$, $k = 1, \dots, K$, with particular emphasis on accurate marginal posterior means and variances.

The key idea is to replace the plug-in updates used by MFVB with conditional moment averaging. For each block $\boldsymbol{\theta}_k$, MP approximates marginal posterior moments by averaging conditional posterior moments of $\boldsymbol{\theta}_k$ over an approximation to the posterior distribution of the

neighbouring blocks. These moments are then matched to a convenient approximating family. This leads to algorithms that retain much of the blockwise structure of MFVB while restoring part of the posterior uncertainty lost through mean-field factorization.

3.1 The basic marginal identity

Let $M(k) \subseteq \{1, \dots, K\} \setminus \{k\}$ denote the indices of the blocks in the Markov blanket of $\boldsymbol{\theta}_k$. By the law of total probability, $p(\boldsymbol{\theta}_k|\mathcal{D}) = \mathbb{E}_{p(\boldsymbol{\theta}_{-k}|\mathcal{D})}[p(\boldsymbol{\theta}_k|\mathcal{D}, \boldsymbol{\theta}_{-k})]$. Using the conditional independence encoded by the Markov blanket, $p(\boldsymbol{\theta}_k|\mathcal{D}, \boldsymbol{\theta}_{-k}) = p(\boldsymbol{\theta}_k|\mathcal{D}, \boldsymbol{\theta}_{M(k)})$, and therefore

$$p(\boldsymbol{\theta}_k|\mathcal{D}) = \mathbb{E}_{p(\boldsymbol{\theta}_{M(k)}|\mathcal{D})}[p(\boldsymbol{\theta}_k|\mathcal{D}, \boldsymbol{\theta}_{M(k)})]. \quad (6)$$

Suppose that $q(\boldsymbol{\theta}_{M(k)}) \approx p(\boldsymbol{\theta}_{M(k)}|\mathcal{D})$. Then (6) suggests the implicit marginal approximation $\tilde{q}_k(\boldsymbol{\theta}_k) = \mathbb{E}_{q(\boldsymbol{\theta}_{M(k)})}[p(\boldsymbol{\theta}_k|\mathcal{D}, \boldsymbol{\theta}_{M(k)})]$. Equivalently, define the local joint approximation

$$r_k(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{M(k)}) = p(\boldsymbol{\theta}_k|\mathcal{D}, \boldsymbol{\theta}_{M(k)})q(\boldsymbol{\theta}_{M(k)}). \quad (7)$$

When both factors on the right-hand side are normalized densities, $r_k(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{M(k)})$ is also normalized, and its marginal density in $\boldsymbol{\theta}_k$ is $\tilde{q}_k(\boldsymbol{\theta}_k)$. In general, direct calculation of $\tilde{q}_k(\boldsymbol{\theta}_k)$ may be difficult. However, its moments are often easier to approximate, e.g.,

$$\mathbb{E}_{\tilde{q}_k}[\mathbf{T}_k(\boldsymbol{\theta}_k)] = \mathbb{E}_{r_k}[\mathbf{T}_k(\boldsymbol{\theta}_k)] = \mathbb{E}_{q(\boldsymbol{\theta}_{M(k)})} \left[\mathbb{E}\{\mathbf{T}_k(\boldsymbol{\theta}_k)|\mathcal{D}, \boldsymbol{\theta}_{M(k)}\} \right]. \quad (8)$$

where $\mathbf{T}_k(\boldsymbol{\theta}_k)$ is a vector of functions. Moment propagation uses (8) to construct a tractable approximation $q_k(\boldsymbol{\theta}_k; \boldsymbol{\eta}_k)$ by choosing $\boldsymbol{\eta}_k$ such that $\mathbb{E}_{q_k(\boldsymbol{\theta}_k; \boldsymbol{\eta}_k)}[\mathbf{T}_k(\boldsymbol{\theta}_k)] = \mathbb{E}_{r_k}[\mathbf{T}_k(\boldsymbol{\theta}_k)]$. Thus, $\tilde{q}_k$ denotes the implicit MP marginal induced by conditional averaging, whereas $q_k(\boldsymbol{\theta}_k; \boldsymbol{\eta}_k)$ denotes a tractable moment-matched approximation to that marginal.

3.2 Moment propagation in conjugate block models

We first consider the case where the conditional density $p(\boldsymbol{\theta}_k|\mathcal{D}, \boldsymbol{\theta}_{M(k)})$ is available in normalized form and its relevant moments are known functions of $\boldsymbol{\theta}_{M(k)}$. In this setting, MP can be

implemented by moment matching without evaluating the full mixture density in (7).

In particular, the posterior mean and covariance of $\boldsymbol{\theta}_k$ can be approximated by

$$\begin{aligned}\mathbb{E}(\boldsymbol{\theta}_k|\mathcal{D}) &\approx \mathbb{E}_{r_k}(\boldsymbol{\theta}_k) = \mathbb{E}_{q(\boldsymbol{\theta}_{M(k)})} \left[\mathbb{E}(\boldsymbol{\theta}_k|\mathcal{D}, \boldsymbol{\theta}_{M(k)}) \right], \quad \text{and} \\ \mathbb{V}(\boldsymbol{\theta}_k|\mathcal{D}) &\approx \mathbb{V}_{r_k}(\boldsymbol{\theta}_k) = \mathbb{E}_{q(\boldsymbol{\theta}_{M(k)})} \left[\mathbb{V}(\boldsymbol{\theta}_k|\mathcal{D}, \boldsymbol{\theta}_{M(k)}) \right] + \mathbb{V}_{q(\boldsymbol{\theta}_{M(k)})} \left[\mathbb{E}(\boldsymbol{\theta}_k|\mathcal{D}, \boldsymbol{\theta}_{M(k)}) \right].\end{aligned}\tag{9}$$

The second term on the right hand side of the expression for $\mathbb{V}_{r_k}(\boldsymbol{\theta}_k)$ in (9) is the additional uncertainty induced by variation in the neighbouring blocks. This is the term that is typically absent in the corresponding MFVB update. Hence (9) provides a direct moment-based correction to the plug-in approximation discussed in Section 2.

More generally, MP can be applied to any vector $\mathbf{T}_k(\boldsymbol{\theta}_k)$ for which the conditional moments $\mathbb{E}[\mathbf{T}_k(\boldsymbol{\theta}_k)|\mathcal{D}, \boldsymbol{\theta}_{M(k)}]$ can be evaluated or approximated. Given these moments, we choose a convenient approximating family $q_k(\boldsymbol{\theta}_k; \boldsymbol{\eta}_k)$ and solve the moment matching equations

$$\mathbb{E}_{q_k(\boldsymbol{\theta}_k; \boldsymbol{\eta}_k)}[\mathbf{T}_k(\boldsymbol{\theta}_k)] = \mathbb{E}_{r_k}[\mathbf{T}_k(\boldsymbol{\theta}_k)].$$

When q_k is an exponential-family density and $\mathbf{T}_k$ is its vector of sufficient statistics, this moment matching step can be interpreted as a KL projection onto that family (Minka, 2005). However, MP is not restricted to exponential-family approximations or to matching sufficient statistics. In later examples, we use multivariate t , inverse-Wishart, Gaussian, and inverse-gamma approximations, and in some cases match moments other than the sufficient statistics when this leads to tractable updates.

3.3 Mean-field moment propagation

Mean-field moment propagation (MFMP) is obtained by approximating the distribution of the neighbouring blocks by a product of their marginal approximations:

$$q(\boldsymbol{\theta}_{M(k)}) = \prod_{j \in M(k)} q_j(\boldsymbol{\theta}_j).$$

This assumption is weaker than the MFVB factorization in the sense that MP updates block k only requires a factorized approximation over the Markov blanket of $\boldsymbol{\theta}_k$, rather than over all parameters. When $K = 2$, no additional factorization is required inside the conditioning block, although the algorithm still constructs separate marginal approximations for the two components. A generic MFMP update for block k has the following form.

1. Given current approximations $\{q_j : j \in M(k)\}$, form $q(\boldsymbol{\theta}_{M(k)}) = \prod_{j \in M(k)} q_j(\boldsymbol{\theta}_j)$.
2. Compute the propagated moments $\mathbf{m}_k^{\text{MP}} = \mathbb{E}_{q(\boldsymbol{\theta}_{M(k)})} \left[\mathbb{E}\{\mathbf{T}_k(\boldsymbol{\theta}_k) | \mathcal{D}, \boldsymbol{\theta}_{M(k)}\} \right]$.
3. Solve for $\boldsymbol{\eta}_k$ such that $\mathbb{E}_{q_k(\boldsymbol{\theta}_k; \boldsymbol{\eta}_k)}[\mathbf{T}_k(\boldsymbol{\theta}_k)] = \mathbf{m}_k^{\text{MP}}$.
4. Replace q_k by the updated density $q_k(\boldsymbol{\theta}_k; \boldsymbol{\eta}_k)$, and cycle through the remaining blocks until convergence.

The following results summarize the asymptotic behaviour of MFMP. Note that the results do not constrain the choice of $\mathcal{Q}_k$ to the exponential family of distributions. Note that we provide only an intuitive condensed statement of each result here. Their full technical statements, required assumptions, and technical proofs are given in Supplementary Section B. The first theorem provide sufficient conditions to guarantee the large-sample convergence of the MFMP solution to the fixed point of a limiting system of integral equations.

Theorem 1 *Assume (A1) to (A5) in Supplementary Section B holds. Then, the MFMP sequence of solutions $\{\hat{\boldsymbol{\lambda}}_{1:K}^{(n)}\}_{n \geq 1}$ converge in probability to the solution of a set of limiting integral equations $\boldsymbol{\lambda}_{1:K}^*$ at rate $O_p(\epsilon_n)$, where ϵ_n is a deterministic sequence according that controls the rate of convergence of the exact posterior.*

The following results establishes the consistency of the MFMP posterior.

Theorem 2 *Assume (A1) to (A6) in Supplementary Section B holds. Then, for each $k = 1, \dots, K$ and any $\epsilon > 0$, $\mathbb{P}_{q_k} \{\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^*\| > \epsilon\} \xrightarrow{\mathbb{P}^*} 0$, where $\mathbb{P}_{q_k}$ is the probability measure corresponding to the k -th marginal MFMP posterior.*

The following results establishes the asymptotic convergence of the MFMP posterior variance.

Theorem 3 Assume (A1) to (A9) holds and $K = 2$. Then, for each $k = 1, 2$, we have

$$\|\Sigma_{kn} - \mathbf{V}_{kn}\| = O_p(n^{-3/2}), \text{ where } \mathbf{V}_{kn} = \mathbb{V}(\boldsymbol{\theta}_k \mid \mathcal{D}_n).$$

Theorem 3 formalizes the main advantage of MFMP over MFVB in the two-component conjugate case: by propagating conditional moment information between blocks, MFMP can recover asymptotically correct marginal posterior variances. The result does not imply that MFMP exactly represents the joint posterior dependence between blocks, but it shows that accurate marginal variance estimation is possible even when the algorithm maintains separate marginal approximations.

MFMP is most useful when the conditional density $p(\boldsymbol{\theta}_k \mid \mathcal{D}, \boldsymbol{\theta}_{M(k)})$ is known and the conditional moments in (8) can be calculated analytically or by low-dimensional numerical integration. This requirement limits its direct application in non-conjugate models. We next introduce a Gaussian alternative that is designed for such settings.

3.4 Gaussian moment propagation

Gaussian moment propagation (GMP) starts from a joint Gaussian approximation $q(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta} - \boldsymbol{\mu}; \Sigma)$, where $\phi(\cdot; \Sigma)$ denotes the multivariate Gaussian density with mean zero and covariance matrix Σ . This Gaussian approximation is used as a propagation device: its conditional distributions are used to construct local marginal approximations, and the first two moments of these local approximations are then used to update the joint Gaussian approximation.

Consider updating block $\boldsymbol{\theta}_k$. After a possible permutation of indices, write

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_k \\ \boldsymbol{\mu}_{-k} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{k,k} & \Sigma_{k,-k} \\ \Sigma_{-k,k} & \Sigma_{-k,-k} \end{bmatrix}.$$

Let $\Sigma_{-k|k} = \Sigma_{-k,-k} - \Sigma_{-k,k} \Sigma_{k,k}^{-1} \Sigma_{k,-k}$. Then the conditional distribution under the current Gaussian approximation is

$$q(\boldsymbol{\theta}_{-k} \mid \boldsymbol{\theta}_k) = \phi\left(\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-k} - \Sigma_{-k,k} \Sigma_{k,k}^{-1} (\boldsymbol{\theta}_k - \boldsymbol{\mu}_k); \Sigma_{-k|k}\right). \quad (10)$$

The GMP update is motivated by the conditional identity from Section 2, i.e., (5). For the

Gaussian conditional distribution in (10), the entropy term $\mathbb{E}_{q(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k)}[-\log q(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k)]$ does not depend on $\boldsymbol{\theta}_k$. If, in addition, the conditional KL term varies slowly as a function of $\boldsymbol{\theta}_k$, then the marginal posterior density of $\boldsymbol{\theta}_k$ is approximated, up to normalization, by

$$r_k(\boldsymbol{\theta}_k) \propto \exp \left[\mathbb{E}_{q(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k)} \{ \log p(\mathcal{D}, \boldsymbol{\theta}) \} \right]. \quad (11)$$

We refer to $r_k(\boldsymbol{\theta}_k)$ as the local GMP marginal approximation for block k . It has dimension d_k , and so can often be normalized and integrated using low-dimensional numerical integration.

Suppose that the first two moments of r_k are available: $\tilde{\boldsymbol{\mu}}_k = \mathbb{E}_{r_k}(\boldsymbol{\theta}_k)$, and $\tilde{\boldsymbol{\Sigma}}_k = \mathbb{V}_{r_k}(\boldsymbol{\theta}_k)$. These moments may be calculated using quadrature, Monte Carlo, Laplace approximation, or other suitable low-dimensional approximations. We then update the joint Gaussian approximation by replacing the marginal distribution of $\boldsymbol{\theta}_k$ with $\tilde{q}_k(\boldsymbol{\theta}_k) = \phi(\boldsymbol{\theta}_k - \tilde{\boldsymbol{\mu}}_k; \tilde{\boldsymbol{\Sigma}}_k)$, while retaining the current conditional distribution $q(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k)$. That is, the updated joint approximation is $\tilde{q}(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k)\tilde{q}_k(\boldsymbol{\theta}_k)$. This is again a multivariate Gaussian distribution, with

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \tilde{\boldsymbol{\mu}}_k \\ \boldsymbol{\mu}_{-k} + \boldsymbol{\Sigma}_{-k,k} \boldsymbol{\Sigma}_{k,k}^{-1} (\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) \end{bmatrix}, \quad \text{and} \quad (12)$$

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_k & \tilde{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_{k,k}^{-1} \boldsymbol{\Sigma}_{k,-k} \\ \boldsymbol{\Sigma}_{-k,k} \boldsymbol{\Sigma}_{k,k}^{-1} \tilde{\boldsymbol{\Sigma}}_k & \boldsymbol{\Sigma}_{-k|-k} + \boldsymbol{\Sigma}_{-k,k} \boldsymbol{\Sigma}_{k,k}^{-1} \tilde{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_{k,k}^{-1} \boldsymbol{\Sigma}_{k,-k} \end{bmatrix}. \quad (13)$$

After replacing $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, the algorithm proceeds to update another block. Cycling through all blocks until convergence yields the GMP approximation.

The GMP algorithm can be summarized as follows.

1. Initialize $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using a Gaussian approximation, such as a Laplace, GVB, MFVB-based Gaussian, or EP approximations.
2. For $k = 1, \dots, K$,
 - Compute the Gaussian conditional distribution $q(\boldsymbol{\theta}_{-k} | \boldsymbol{\theta}_k)$ from (11).
 - Form the local marginal approximation $r_k(\boldsymbol{\theta}_k)$ in (11).
 - Compute $\tilde{\boldsymbol{\mu}}_k = \mathbb{E}_{r_k}(\boldsymbol{\theta}_k)$, and $\tilde{\boldsymbol{\Sigma}}_k = \mathbb{V}_{r_k}(\boldsymbol{\theta}_k)$.

- Update the joint Gaussian approximation using (12) and (13).
3. Repeat until convergence.

The local density $r_k(\boldsymbol{\theta}_k)$ and the Gaussian density $q(\boldsymbol{\theta})$ play different roles. The Gaussian density $q(\boldsymbol{\theta})$ is used to propagate dependence information between blocks through its conditional distributions. The local density $r_k(\boldsymbol{\theta}_k)$ can be used directly as the marginal posterior approximation for $\boldsymbol{\theta}_k$, while its first two moments are used to update $q(\boldsymbol{\theta})$.

Theorem 4 *If $\boldsymbol{\Sigma}$ and $\tilde{\boldsymbol{\Sigma}}_k$ are positive definite, then the updated covariance matrix $\tilde{\boldsymbol{\Sigma}}$ in (13) is positive definite.*

The proof is given in Supplementary Section C.3. Theorem 4 ensures that each GMP update preserves positive definiteness of the Gaussian propagation covariance matrix. Notice that the updated off-diagonal blocks are nonzero if and only if the corresponding off-diagonal blocks in the current $\boldsymbol{\Sigma}$ are nonzero. Therefore, GMP should be initialized with a Gaussian approximation that already contains useful posterior dependence information. In practice, we have found that the quality of the final GMP approximation depends on the quality of this initialization.

3.5 Hybrid and approximate moment propagation updates

The preceding sections describe two idealized MP updates. MFMP is attractive when conditional posterior moments are available, while GMP requires a reasonable joint Gaussian approximation. In practice, neither set of requirements may hold exactly. The required moments in (8), the expectations in (11), or the moment matching equations for $\boldsymbol{\eta}_k$ may be analytically intractable. In such cases, MP can be implemented using hybrid or approximate updates.

One possibility is to approximate the required moments using Monte Carlo, stochastic approximation, quadrature, Laplace approximation, or delta-method expansions. Another possibility is to apply MP only to blocks for which the conditional moment calculations are tractable, and use standard VB-style updates for the remaining blocks. For example, suppose $V \subset M(k)$ is a subset of neighbouring blocks over which exact conditional moment propagation is inconvenient. Define $\tilde{p}_k(\boldsymbol{\theta}_k | \mathcal{D}, \boldsymbol{\theta}_{M(k) \setminus V}) \propto \exp \left[\mathbb{E}_{q(\boldsymbol{\theta}_V)} \{ \log p(\boldsymbol{\theta}_k | \mathcal{D}, \boldsymbol{\theta}_{M(k)}) \} \right]$, where

$q(\boldsymbol{\theta}_V) = \prod_{j \in V} q_j(\boldsymbol{\theta}_j)$. If $\tilde{p}_k$ can be normalized and its moments computed, then the MP moment update can be replaced by

$$\mathbb{E}[\mathbf{T}_k(\boldsymbol{\theta}_k) | \mathcal{D}] \approx \mathbb{E}_{q(\boldsymbol{\theta}_{M(k) \setminus V})} \left[\mathbb{E}_{\tilde{p}_k(\boldsymbol{\theta}_k | \mathcal{D}, \boldsymbol{\theta}_{M(k) \setminus V})} \{ \mathbf{T}_k(\boldsymbol{\theta}_k) \} \right]. \quad (14)$$

This is a hybrid MP/VB update: the blocks in V are averaged in a variational manner, while the remaining neighbouring blocks are propagated through conditional moments.

A second hybrid strategy is useful when a full conditional approximation $q(\boldsymbol{\theta}_{-k} | \boldsymbol{\theta}_k)$ is unavailable, but a lower-dimensional structured approximation is available. Suppose, for example, that for some $k' \neq k$, $q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k'}) \prod_{j \notin \{k, k'\}} q_j(\boldsymbol{\theta}_j)$, and that $q(\boldsymbol{\theta}_{k'} | \boldsymbol{\theta}_k)$ is available. Then a lower-dimensional replacement for the GMP update is

$$r_k(\boldsymbol{\theta}_k) \propto \exp \left[\mathbb{E}_{q(\boldsymbol{\theta}_{k'} | \boldsymbol{\theta}_k) \prod_{j \notin \{k, k'\}} q_j(\boldsymbol{\theta}_j)} \{ \log p(\mathcal{D}, \boldsymbol{\theta}) \} \right]. \quad (15)$$

This approximation propagates dependence between $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k'}$, while treating the remaining blocks in a mean-field manner. Hybrid updates such as (14) and (15) may be less accurate than the ideal MFMP or GMP updates, but they can substantially expand the range of models for which moment propagation is computationally feasible. In general, MFMP is most useful when conditional posterior moments are available analytically or through low-dimensional integration; GMP is most useful when a good joint Gaussian approximation is available; and hybrid updates are useful when only part of the model admits tractable MP updates.

4 Examples

In this section, we illustrate moment propagation through a sequence of examples. The examples are chosen to demonstrate different aspects of the methodology. Linear regression and the multivariate normal model provide benchmark settings where exact marginal posterior distributions are available, allowing us to verify that suitable MP updates can recover the exact answer. The multivariate normal model with missing data demonstrates how MP can propagate uncertainty from latent quantities. Probit regression illustrates the use of MP in an auxiliary-variable

model, including a case where not all approximating densities need to be specified explicitly. Generalized linear models demonstrate Gaussian moment propagation (GMP) in non-conjugate settings. Finally, the Bayesian Lasso example shows how GMP can be adapted to a model with a non-smooth prior. Detailed derivations, theory, and algorithmic summaries are provided in the Supplementary Material.

4.1 Linear regression

We begin with a conjugate linear regression model. This example is useful because the exact marginal posterior distributions are available in closed form, making it possible to assess whether moment propagation can recover the exact posterior.

Consider the linear model $\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, with priors $\boldsymbol{\beta}|\sigma^2 \sim N_p(0, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$, $\sigma^2 \sim \text{IG}(A, B)$, where $g > 0$ is fixed. Let $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, and $u = g/(1 + g)$. The posterior distribution is available in closed form. We consider two MP approximations. The first uses a Gaussian density for $q(\boldsymbol{\beta})$ and shows how MP improves upon the usual MFVB variance update. The second uses a multivariate t density for $q(\boldsymbol{\beta})$ and recovers the exact marginal posterior.

4.1.1 Gaussian approximation for $q(\boldsymbol{\beta})$

First suppose $q(\boldsymbol{\beta}) = N_p(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, and $q(\sigma^2) = \text{IG}(\tilde{A}, \tilde{B})$. To update $q(\boldsymbol{\beta})$ given $q(\sigma^2)$, we propagate the conditional moments of $\boldsymbol{\beta}|\mathbf{y}, \sigma^2$. This gives

$$\mathbb{E}_r(\boldsymbol{\beta}) = u\hat{\boldsymbol{\beta}}, \quad \text{and} \quad \mathbb{V}_r(\boldsymbol{\beta}) = \frac{\tilde{B}}{\tilde{A} - 1}u(\mathbf{X}^T\mathbf{X})^{-1}.$$

Matching these moments with those of $q(\boldsymbol{\beta})$ gives leads to the updates $\tilde{\boldsymbol{\mu}} \leftarrow u\hat{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{\Sigma}} \leftarrow (\tilde{B}/(\tilde{A} - 1))u(\mathbf{X}^T\mathbf{X})^{-1}$. The update for $q(\sigma^2)$, uses the conditional distribution $\sigma^2|\mathbf{y}, \boldsymbol{\beta} \sim \text{IG}(A_{n,p}, \mathcal{B}(\boldsymbol{\beta}))$, where $A_{n,p} = A + (n + p)/2$ and $\mathcal{B}(\boldsymbol{\beta}) = B + (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2)/2 + (2g)^{-1}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$.

Moment propagation gives

$$\mathbb{E}_r(\sigma^2) = \frac{\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})]}{A_{n,p} - 1}, \quad \text{and} \quad \mathbb{V}_r(\sigma^2) = \frac{\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})]^2}{(A_{n,p} - 1)^2(A_{n,p} - 2)} + \frac{\mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})]}{(A_{n,p} - 1)(A_{n,p} - 2)}.$$

Here $\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})] = B + \|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\mu}}\|^2/2 + \tilde{\boldsymbol{\mu}}^T\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\mu}}/(2g) + \text{tr}(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\Sigma}})/(2u)$, and $\mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})] =$

$\text{tr}[(\mathbf{X}^T \mathbf{X} \tilde{\Sigma})^2]/(2u^2)$. Note that the MFVB approximation leads to the same approximation of posterior mean, and equals $\mathbb{V}_r(\sigma^2)$ without the additional second term. Matching the mean and variance of $q(\sigma^2)$ to $\mathbb{E}_r(\sigma^2)$ and $\mathbb{V}_r(\sigma^2)$ leads to the updates

$$\tilde{A} \leftarrow \frac{\mathbb{E}_r(\sigma^2)^2}{\mathbb{V}_r(\sigma^2)} + 2, \quad \tilde{B} \leftarrow (\tilde{A} - 1)\mathbb{E}_r(\sigma^2).$$

This approximation shows the key MP correction: the variance of $B(\beta)$ contributes an additional term to the variance update for σ^2 . However, $q(\beta)$ is still misspecified because the exact marginal posterior of β is multivariate t . The next approach corrects this.

4.1.2 Multivariate t approximation for $q(\beta)$

Now suppose $q(\beta) = t_p(\tilde{\mu}, \tilde{\Sigma}, \tilde{\nu})$, and $q(\sigma^2) = \text{IG}(\tilde{A}, \tilde{B})$. For fixed $q(\sigma^2)$, we match the first two moments of β , and $Q(\beta)^2$, where $Q(\beta) = (\beta - \tilde{\mu})^T \tilde{\Sigma}^{-1} (\beta - \tilde{\mu})$. This gives

$$\tilde{\mu} \leftarrow u\hat{\beta}, \quad \tilde{\Sigma} \leftarrow \frac{\tilde{B}}{\tilde{A}} u(\mathbf{X}^T \mathbf{X})^{-1}, \quad \tilde{\nu} \leftarrow 2\tilde{A}.$$

The update for $q(\sigma^2)$ again matches $\mathbb{E}_r(\sigma^2)$ and $\mathbb{V}_r(\sigma^2)$, but now the moments of $B(\beta)$ are computed under the multivariate t approximation:

$$\begin{aligned} \mathbb{E}_q[B(\beta)] &= B + \frac{1}{2} \|\mathbf{y} - \mathbf{X}\tilde{\mu}\|^2 + \frac{1}{2g} \tilde{\mu}^T \mathbf{X}^T \mathbf{X} \tilde{\mu} + \frac{\tilde{\nu}}{\tilde{\nu} - 2} \frac{\text{tr}(\mathbf{X}^T \mathbf{X} \tilde{\Sigma})}{2u} \quad \text{and} \\ \mathbb{V}_q[B(\beta)] &= \frac{1}{2u^2} \left[\frac{\tilde{\nu}^2 \text{tr}\{(\mathbf{X}^T \mathbf{X} \tilde{\Sigma})^2\}}{(\tilde{\nu} - 2)(\tilde{\nu} - 4)} + \frac{\tilde{\nu}^2 \{\text{tr}(\mathbf{X}^T \mathbf{X} \tilde{\Sigma})\}^2}{(\tilde{\nu} - 2)^2(\tilde{\nu} - 4)} \right], \end{aligned}$$

provided the required moments exist. Matching the mean and variance of $q(\sigma^2)$ gives the same inverse-gamma moment equations as above. The Supplementary Material [D](#) shows that this second MP approach recovers the exact marginal posterior distributions $p(\beta|\mathbf{y})$ and $p(\sigma^2|\mathbf{y})$. Thus, the example demonstrates that MP can recover exact posterior marginals when the propagated moments are exact and the approximating families are sufficiently rich.

4.2 Multivariate normal model

The next example gives a second exact-recovery benchmark. Let $\mathbf{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, \dots, n$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathcal{S}_p^+$. We use the conditionally conjugate priors $\boldsymbol{\mu} | \boldsymbol{\Sigma} \sim N_p(0, \lambda_0^{-1} \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma} \sim \text{IW}_p(\boldsymbol{\Psi}_0, \nu_0)$, an inverse-Wishart distribution, where $\lambda_0 > 0$, $\nu_0 > p - 1$, and $\boldsymbol{\Psi}_0 \in \mathcal{S}_p^+$. Without loss of generality, the prior mean of $\boldsymbol{\mu}$ is taken to be zero, after translating the observations if necessary. This model has a closed-form posterior distribution.

We approximate the marginal posterior distributions by $q(\boldsymbol{\mu}) = t_p(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \tilde{\nu})$, and $q(\boldsymbol{\Sigma}) = \text{IW}_p(\tilde{\boldsymbol{\Psi}}, \tilde{d})$. The MP update for $q(\boldsymbol{\mu})$ matches $\mathbb{E}_r(\boldsymbol{\mu})$, $\mathbb{V}_r(\boldsymbol{\mu})$, and a fourth radial moment to give

$$\tilde{\boldsymbol{\mu}} \leftarrow \frac{n\bar{\mathbf{x}}}{\lambda_n}, \quad \tilde{\boldsymbol{\Sigma}} \leftarrow \frac{\tilde{\boldsymbol{\Psi}}}{\lambda_n(\tilde{d} - p + 1)}, \quad \tilde{\nu} \leftarrow \tilde{d} - p + 1,$$

where $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i$ and $\lambda_n = \lambda_0 + n$.

To update $q(\boldsymbol{\Sigma})$, we match $\mathbb{E}_r(\boldsymbol{\Sigma})$ and the diagonal element-wise variances of $\boldsymbol{\Sigma}$. Let $\boldsymbol{\Psi}_n = \boldsymbol{\Psi}_0 + \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T - n\bar{\mathbf{x}} \bar{\mathbf{x}}^T + (\lambda_0/\lambda_n)\bar{\mathbf{x}} \bar{\mathbf{x}}^T$, and $\nu_n = \nu_0 + n$. The required propagated moments can be written in the form

$$\begin{aligned} \mathbb{E}_r(\boldsymbol{\Sigma}) &= \frac{\mathbf{A}}{\nu_n - p}, \quad \text{and} \quad \text{dg}[\mathbb{E}\mathbb{W}\mathbb{V}_r(\boldsymbol{\Sigma})] = \frac{2\text{dg}(\mathbf{A})^{\odot 2} + (\nu_n - p)\text{dg}(\mathbf{B})}{(\nu_n - p)^2(\nu_n - p - 2)}, \\ \text{where} \quad \mathbf{A} &= \boldsymbol{\Psi}_n + \frac{\lambda_n \tilde{\nu}}{\tilde{\nu} - 2} \tilde{\boldsymbol{\Sigma}}, \quad \text{and} \quad \text{dg}(\mathbf{B}) = \frac{2\lambda_n^2 \tilde{\nu}^2 (\tilde{\nu} - 1)}{(\tilde{\nu} - 2)^2 (\tilde{\nu} - 4)} \text{dg}(\tilde{\boldsymbol{\Sigma}})^{\odot 2}. \end{aligned}$$

Here $\text{dg}(\mathbf{A})$ denotes the vector of diagonal elements of a square matrix $\mathbf{A}$, and $\mathbb{E}\mathbb{W}\mathbb{V}(\mathbf{A})$ denotes the matrix of element-wise variances of $\mathbf{A}$. Matching these quantities to the corresponding moments of an inverse-Wishart distribution gives

$$\tilde{d} \leftarrow \frac{2\mathbf{1}_p^T [\text{dg}(\mathbb{E}_r(\boldsymbol{\Sigma}))^{\odot 2}]}{\mathbf{1}_p^T \text{dg}[\mathbb{E}\mathbb{W}\mathbb{V}_r(\boldsymbol{\Sigma})]} + p + 3, \quad \tilde{\boldsymbol{\Psi}} \leftarrow (\tilde{d} - p - 1) \mathbb{E}_r(\boldsymbol{\Sigma}).$$

The Supplementary Material [E](#) gives the full derivation and shows that the MP algorithm has a fixed point corresponding to the exact marginal posteriors under assumptions stated there.

4.3 Multivariate normal model with missing data

We now extend the multivariate normal example to the case where some components of the observations are missing. This example illustrates how MFMP can propagate uncertainty from latent variables into the posterior approximation for model parameters.

Assume the model from Section 4.2, but suppose that for each observation $\mathbf{x}_i$, only a subset of its components is observed. We assume an ignorable missingness mechanism, so that the missingness process does not enter the posterior distribution of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, or the missing data.

Let O_i and M_i denote the observed and missing index sets for observation i , respectively. Let $\mathbf{z}_i = \mathbf{x}_{i,M_i}$ be the vector of missing components. Define $\mathbf{c}_i$ to be the vector obtained from $\mathbf{x}_i$ by replacing missing entries with zeros, and let $\mathbf{R}_i$ be the matrix that inserts $\mathbf{z}_i$ into the missing positions. Then $\mathbf{x}_i = \mathbf{c}_i + \mathbf{R}_i \mathbf{z}_i$. Let $\mathbf{X}_M$ denote the collection of missing components.

We use the MFMP factorization $q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X}_M) = q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i \in \mathcal{M}} q(\mathbf{z}_i)$, where $\mathcal{M}$ indexes observations with at least one missing component, and $q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \text{NIW}_p(\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\boldsymbol{\Psi}}, \tilde{\nu})$, the p -dimensional Normal–Inverse–Wishart distribution, and $q(\mathbf{z}_i) = N_{|M_i|}(\tilde{\mathbf{m}}_i, \tilde{\mathbf{V}}_i)$, for $i \in \mathcal{M}$. The update for each missing vector $\mathbf{z}_i$ is obtained by averaging the Gaussian conditional distribution of the missing components given the observed components over the current approximation to $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The update for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then treats the completed data matrix through its first two moments, thereby propagating missing-data uncertainty into the parameter approximation.

For each $i \in \mathcal{M}$, moment matching gives $\tilde{\mathbf{m}}_i \leftarrow \mathbb{E}_r(\mathbf{z}_i)$, and $\tilde{\mathbf{V}}_i \leftarrow \mathbb{V}_r(\mathbf{z}_i)$. If $O_i = \emptyset$ then $\mathbb{E}_r(\mathbf{z}_i) = \tilde{\boldsymbol{\mu}}$ and $\mathbb{V}_r(\mathbf{z}_i) = (1 + \tilde{\lambda}^{-1})\tilde{\boldsymbol{\Psi}}/(\tilde{\nu} - p - 1)$ otherwise $\mathbb{E}_r(\mathbf{z}_i) = \mathbf{L}_i(\mathbf{x}_{i,O_i} - \tilde{\boldsymbol{\mu}}_{O_i})$

$$\text{and} \quad \mathbb{V}_r(\mathbf{z}_i) = \left[1 + \tilde{\lambda}^{-1} + (\mathbf{x}_{i,O_i} - \tilde{\boldsymbol{\mu}}_{O_i})^T \tilde{\boldsymbol{\Psi}}_{O_i O_i}^{-1} (\mathbf{x}_{i,O_i} - \tilde{\boldsymbol{\mu}}_{O_i}) \right] \frac{\tilde{\boldsymbol{\Psi}}_{M_i | O_i}}{\tilde{\nu} - |M_i| - 1},$$

where $\mathbf{L}_i = \mathbb{E}_q(\boldsymbol{\Sigma}_{M_i O_i} \boldsymbol{\Sigma}_{O_i O_i}^{-1}) = \tilde{\boldsymbol{\Psi}}_{M_i O_i} \tilde{\boldsymbol{\Psi}}_{O_i O_i}^{-1}$, and $\tilde{\boldsymbol{\Psi}}_{M_i | O_i} = \tilde{\boldsymbol{\Psi}}_{M_i M_i} - \tilde{\boldsymbol{\Psi}}_{M_i O_i} \tilde{\boldsymbol{\Psi}}_{O_i O_i}^{-1} \tilde{\boldsymbol{\Psi}}_{O_i M_i}$.

For the update of $q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, define $\hat{\mathbf{x}}_i = \mathbb{E}_q(\mathbf{x}_i) = \mathbf{c}_i + \mathbf{R}_i \tilde{\mathbf{m}}_i$, $\mathbf{B}_i = \mathbb{V}_q(\mathbf{x}_i) = \mathbf{R}_i \tilde{\mathbf{V}}_i \mathbf{R}_i^T$, $\hat{\bar{\mathbf{x}}} = \sum_{i=1}^n \hat{\mathbf{x}}_i / n$, $\lambda_n = \lambda_0 + n$, and $\mathbf{A} = \sum_{i=1}^n \mathbf{B}_i$. Then

$$\mathbb{E}_r(\boldsymbol{\mu}) = \frac{n \hat{\bar{\mathbf{x}}}}{\lambda_n} \equiv \hat{\boldsymbol{\mu}}_n, \quad \text{and} \quad \mathbb{V}_r(\boldsymbol{\mu}) = \frac{\mathbb{E}_q(\boldsymbol{\Sigma})}{\lambda_n} + \frac{\mathbf{A}}{\lambda_n^2}.$$

Similarly, $\mathbb{E}_r(\boldsymbol{\Sigma}) = \mathbf{C}/(\nu_n - p - 1)$, where $\mathbf{C} = \mathbb{E}_q(\boldsymbol{\Psi}_n) = \boldsymbol{\Psi}_0 + \sum_{i=1}^n \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^T + (1 - \lambda_n^{-1})\mathbf{A} - \lambda_n \hat{\boldsymbol{\mu}}_n \hat{\boldsymbol{\mu}}_n^T$.

The diagonal element-wise variances can be computed from the moments of the completed data; the resulting formulae are given in Supplementary Material F. Matching $\mathbb{E}_r(\boldsymbol{\mu})$, $\text{tr}[\mathbb{V}_r(\boldsymbol{\mu})]$, $\mathbb{E}_r(\boldsymbol{\Sigma})$, and $\mathbf{1}_p^T \text{dg}[\mathbb{E}\mathbb{W}\mathbb{V}_r(\boldsymbol{\Sigma})]$ to the corresponding NIW moments gives

$$\tilde{\boldsymbol{\mu}} \leftarrow \mathbb{E}_r(\boldsymbol{\mu}), \quad \tilde{\lambda} \leftarrow \frac{\text{tr}[\mathbb{E}_r(\boldsymbol{\Sigma})]}{\text{tr}[\mathbb{V}_r(\boldsymbol{\mu})]}, \quad \tilde{\nu} \leftarrow \frac{2\mathbf{1}_p^T [\text{dg}\{\mathbb{E}_r(\boldsymbol{\Sigma})\}]^{\odot 2}}{\mathbf{1}_p^T \text{dg}[\mathbb{E}\mathbb{W}\mathbb{V}_r(\boldsymbol{\Sigma})]} + p + 3, \quad \tilde{\boldsymbol{\Psi}} \leftarrow (\tilde{\nu} - p - 1)\mathbb{E}_r(\boldsymbol{\Sigma}).$$

4.4 Probit regression

We next consider probit regression. This example illustrates a useful feature of MP: not every auxiliary quantity requires a fully specified approximating density. In this case, the update for $q(\boldsymbol{\beta})$ requires only the first two moments of the auxiliary variables, so MP can propagate those moments directly. Let $y_i \in \{0, 1\}$, and let $\mathbf{x}_i \in \mathbb{R}^p$. The probit likelihood can be written as

$$p(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^n \Phi(\mathbf{x}_i^T \boldsymbol{\beta})^{y_i} [1 - \Phi(\mathbf{x}_i^T \boldsymbol{\beta})]^{1-y_i} = \prod_{i=1}^n \Phi(\mathbf{z}_i^T \boldsymbol{\beta}), \quad (16)$$

where $\mathbf{z}_i = (2y_i - 1)\mathbf{x}_i$. Assume the Gaussian prior $\boldsymbol{\beta} \sim N_p(\mathbf{0}, \mathbf{D}^{-1})$, where $\mathbf{D}$ is positive definite. Using the Albert–Chib auxiliary-variable representation (Albert and Chib, 1993), $p(y_i, a_i|\boldsymbol{\beta}) = \phi(a_i - \mathbf{z}_i^T \boldsymbol{\beta})\mathbb{I}(a_i > 0)$, $i = 1, \dots, n$. Integrating over a_i recovers $p(y_i|\boldsymbol{\beta}) = \Phi(\mathbf{z}_i^T \boldsymbol{\beta})$. Let $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{Z} = \text{diag}(2\mathbf{y} - \mathbf{1}_n)\mathbf{X}$.

We use $q(\boldsymbol{\beta}) = N_p(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, but do not specify a full parametric form for $q(\mathbf{a})$. Instead, at least at a conceptual level, we maintain only $\mathbb{E}_q(\mathbf{a})$ and $\mathbb{V}_q(\mathbf{a})$ because these are the quantities needed to update $q(\boldsymbol{\beta})$. Later, we merge updates so that $\mathbb{V}_q(\mathbf{a})$ is not explicitly stored. Matching the propagated moments of $\boldsymbol{\beta}$ gives $\tilde{\boldsymbol{\mu}} \leftarrow \mathbf{S}\mathbf{Z}^T \mathbb{E}_q(\mathbf{a})$, and $\tilde{\boldsymbol{\Sigma}} \leftarrow \mathbf{S} + \mathbf{S}\mathbf{Z}^T \mathbb{V}_q(\mathbf{a})\mathbf{Z}\mathbf{S}$, where $\mathbf{S} = (\mathbf{Z}^T \mathbf{Z} + \mathbf{D})^{-1}$. The required moments of $\mathbf{a}$ are obtained from the truncated-normal conditional distribution $\mathbf{a}|\mathbf{y}, \boldsymbol{\beta}$. Let $\zeta_k(t) = d^k \log \Phi(t)/dt^k$. Then

$$\mathbb{E}_q(\mathbf{a}) \leftarrow \mathbf{Z}\tilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}} + \mathbb{E}_q[\zeta_1(\mathbf{Z}\boldsymbol{\beta})], \quad \text{and} \quad \mathbb{V}_q(\mathbf{a}) \leftarrow \mathbf{I}_n + \text{diag}[\mathbb{E}_q\{\zeta_2(\mathbf{Z}\boldsymbol{\beta})\}] + \mathbb{V}_q[\mathbf{Z}\boldsymbol{\beta} + \zeta_1(\mathbf{Z}\boldsymbol{\beta})].$$

The expectations $\mathbb{E}_q[\zeta_1(\mathbf{Z}\boldsymbol{\beta})]$ and $\mathbb{E}_q[\zeta_2(\mathbf{Z}\boldsymbol{\beta})]$ can be evaluated element-wise via one dimensional quadrature. The full covariance term $\mathbb{V}_q[\mathbf{Z}\boldsymbol{\beta} + \zeta_1(\mathbf{Z}\boldsymbol{\beta})]$ would require two-dimensional

quadrature for each pair of observations. For computational efficiency, we instead use a first-order delta-method approximation:

$$\mathbb{V}_q(\mathbf{a}) \approx \mathbf{I}_n + \text{diag} [\mathbb{E}_q\{\zeta_2(\mathbf{Z}\boldsymbol{\beta})\}] + [\mathbf{I}_n + \text{diag}\{\zeta_2(\mathbf{Z}\tilde{\boldsymbol{\mu}})\}]\mathbf{Z}\tilde{\boldsymbol{\Sigma}}\mathbf{Z}^T[\mathbf{I}_n + \text{diag}\{\zeta_2(\mathbf{Z}\tilde{\boldsymbol{\mu}})\}].$$

Substituting this expression directly into the update for $\tilde{\boldsymbol{\Sigma}}$ avoids storing the full $n \times n$ covariance matrix $\mathbb{V}_q(\mathbf{a})$. The resulting combined updates are

$$\begin{aligned} \mathbb{E}_q(\mathbf{a}) &\leftarrow \mathbf{Z}\tilde{\boldsymbol{\mu}} + \mathbb{E}_q[\zeta_1(\mathbf{Z}\boldsymbol{\beta})], & \tilde{\boldsymbol{\mu}} &\leftarrow \mathbf{S}\mathbf{Z}^\top \mathbb{E}_q(\mathbf{a}), & \text{and} \\ \tilde{\boldsymbol{\Sigma}} &\leftarrow \mathbf{S} + \mathbf{S}\mathbf{Z}^T[\mathbf{I}_n + \text{diag}\{\mathbb{E}_q(\zeta_2(\mathbf{Z}\boldsymbol{\beta}))\}]\mathbf{Z}\mathbf{S} \\ &\quad + \mathbf{S}\mathbf{Z}^T[\mathbf{I}_n + \text{diag}\{\zeta_2(\mathbf{Z}\tilde{\boldsymbol{\mu}})\}]\mathbf{Z}\tilde{\boldsymbol{\Sigma}}\mathbf{Z}^T[\mathbf{I}_n + \text{diag}\{\zeta_2(\mathbf{Z}\tilde{\boldsymbol{\mu}})\}]\mathbf{Z}\mathbf{S}. \end{aligned}$$

A careful implementation costs $O(np^2 + p^3)$ per iteration. The following result summarizes the large-sample behaviour of the fixed point. The proof is given in Supplementary Material [G](#).

Theorem 5 *Under the regularity conditions stated in the Supplementary Material Section [G.3](#), the fixed point of the probit MP updates satisfies*

$$\|\tilde{\boldsymbol{\mu}}_\beta^* - \hat{\boldsymbol{\beta}}\| = O_p(n^{-1}), \quad \|\tilde{\boldsymbol{\Sigma}}_\beta^* - \mathbf{V}\|_F = O_p(n^{-2}),$$

where $\hat{\boldsymbol{\beta}}$ is the posterior mode and $\mathbf{V} = [\mathbf{Z}^T \text{diag}\{-\zeta_2(\mathbf{Z}\hat{\boldsymbol{\beta}})\}\mathbf{Z} + \mathbf{D}]^{-1}$.

Thus, the MP covariance approximation is asymptotically equivalent to the usual curvature-based covariance approximation. The result does not follow directly from Theorem [1](#) because of the first-order delta-method approximation used above.

4.5 GMP for generalized linear models

We now consider non-conjugate generalized linear models. This example illustrates the GMP construction from Section [3.4](#): a joint Gaussian approximation is used as a propagation device, while low-dimensional local marginal approximations are used for inference and updating.

Consider a one-parameter exponential-family likelihood with Gaussian prior

$$p(\mathbf{y}|\boldsymbol{\theta}) = \exp[\mathbf{1}_n^T f(\mathbf{y}, \mathbf{X}\boldsymbol{\theta})], \quad \boldsymbol{\theta} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_p), \quad \text{where } f(\mathbf{y}, \mathbf{X}\boldsymbol{\theta}) = (f(y_1, \mathbf{x}_1^T \boldsymbol{\theta}), \dots, f(y_n, \mathbf{x}_n^T \boldsymbol{\theta})).$$

Examples include logistic regression, with $f(y, x) = yx - \log(1 + \exp x)$, probit regression, with $f(y, x) = \log \Phi[(2y - 1)x]$, and Poisson regression, with $f(y, x) = yx - \exp(x) - \log(y!)$.

Suppose that an initial Gaussian approximation $q(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta} - \boldsymbol{\mu}; \boldsymbol{\Sigma})$ has been obtained, for example from a Laplace approximation, GVB, or EP. We use the component-wise partition $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_p\}$. For the update of θ_k , the Gaussian conditional distribution gives $q(\boldsymbol{\theta}_{-k}|\theta_k) = \phi(\mathbf{a}_k \theta_k + \mathbf{b}_k, \boldsymbol{\Sigma}_{-k|k})$, where $\mathbf{a}_k = \boldsymbol{\Sigma}_{-k,k}/\Sigma_{k,k}$, $\mathbf{b}_k = \boldsymbol{\mu}_{-k} - \mathbf{a}_k \mu_k$, and $\boldsymbol{\Sigma}_{-k|k} = \boldsymbol{\Sigma}_{-k,-k} - \boldsymbol{\Sigma}_{-k,k} \Sigma_{k,k}^{-1} \boldsymbol{\Sigma}_{k,-k}$. It follows that $\mathbf{X} \mathbb{E}_q(\boldsymbol{\theta}|\theta_k) = \mathbf{c}_k \theta_k + \mathbf{d}_k$, where $\mathbf{c}_k = \mathbf{X}_k + \mathbf{X}_{-k} \mathbf{a}_k$, and $\mathbf{d}_k = \mathbf{X}_{-k} \mathbf{b}_k$. Let $\mathbf{s}_{-k|k}^2 = \text{dg}(\mathbf{X}_{-k} \boldsymbol{\Sigma}_{-k|k} \mathbf{X}_{-k}^T)$, and define $F(y, \mu, \sigma^2) = \int_{-\infty}^{\infty} f(y, x) \phi(x - \mu; \sigma^2) dx$. Then the GMP local marginal approximation from Section 3.4 is

$$r_k(\theta_k) \propto \exp \left[\mathbf{1}_n^T F(\mathbf{y}, \mathbf{c}_k \theta_k + \mathbf{d}_k, \mathbf{s}_{-k|k}^2) - \frac{\|\mathbf{a}_k\|^2 + 1}{2\sigma^2} \theta_k^2 - \frac{\mathbf{a}_k^T \mathbf{b}_k}{\sigma^2} \theta_k \right]. \quad (17)$$

The only model-specific ingredient is the function F , the Gaussian average of the scalar log-likelihood contribution. For Poisson regression, $F(y, \mu, \sigma^2) = y\mu - \exp(\mu + \sigma^2/2) - \log(y!)$, whereas for logistic and probit regression it is evaluated using numerical quadrature.

After normalizing r_k , we compute $\tilde{\mu}_k = \mathbb{E}_{r_k}(\theta_k)$, and $\tilde{\Sigma}_k = \mathbb{V}_{r_k}(\theta_k)$, and update the Gaussian propagation approximation using the GMP update equations from Section 3.4. The density r_k itself can be retained as the marginal approximation for θ_k .

For a single coordinate update, the deterministic algebraic part of the GMP update can be implemented in $O(np + p^2)$ time (see Supplementary Material C), and a full sweep over $k = 1, \dots, p$ has cost $O(np^2 + p^3)$. If N denotes the number of quadrature nodes then evaluating (17) requires approximating F per sample costing $O(Nn)$, and then normalizing r_k costing an additional $O(N)$ per coordinate. Thus the cost of quadrature is $O(Nnp)$ per iteration. For the Poisson case this is reduced to $O(Np)$.

4.6 GMP for Bayesian Lasso

The final example considers a Bayesian Lasso model. This example illustrates a hybrid GM-P/VB update: dependence among regression coefficients is propagated through a Gaussian approximation to $q(\boldsymbol{\beta})$, while uncertainty in σ^2 is averaged using a variational factor $q(\sigma^2)$. It also represents an example where the prior is not smooth.

Consider the linear model $\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$, with $\boldsymbol{\beta}|\sigma^2 \sim \text{Laplace}(0, \sigma/\lambda)$ and $\sigma^2 \sim \text{IG}(A, B)$, where $\lambda > 0$ is a fixed constant. We assume that $\mathbf{y}$ and $\mathbf{X}$ have been centered so that the intercept can be omitted. The dependence of the Laplace prior on σ^2 is useful because, for fixed λ , the resulting joint posterior is unimodal. The joint density is proportional to

$$p(\mathcal{D}, \boldsymbol{\beta}, \sigma^2) \propto \exp \left[- \left(A + \frac{n}{2} + p + 1 \right) \log \sigma^2 - \frac{B}{\sigma^2} - \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 - \frac{\lambda}{\sigma} \|\boldsymbol{\beta}\|_1 \right].$$

For fixed σ^2 , the negative log posterior is proportional in $\boldsymbol{\beta}$ to the usual Lasso objective, up to the prior scaling. Suppose an initial approximation has been obtained of the form $q(\boldsymbol{\beta}, \sigma^2) = q(\boldsymbol{\beta})q(\sigma^2)$, where $q(\boldsymbol{\beta}) = \phi(\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, and $q(\sigma^2) = \text{IG}(\tilde{A}, \tilde{B})$. For the update of $\boldsymbol{\beta}_k$, define $\mathbf{a}_k$, $\mathbf{b}_k$, $\mathbf{c}_k$, $\mathbf{d}_k$, $\boldsymbol{\Sigma}_{-k|k}$ as in Section 4.5. Using the hybrid GMP/VB approximation,

$$\begin{aligned} r_k(\beta_k) &\propto \exp \left[\mathbb{E}_{q(\boldsymbol{\beta}_{-k}|\beta_k)q(\sigma^2)} \{ \log p(\mathcal{D}, \boldsymbol{\beta}, \sigma^2) \} \right] \\ &\propto \exp \left[- \frac{\tilde{A}}{2\tilde{B}} \left\{ \|\mathbf{c}_k\|^2 \beta_k^2 - 2\mathbf{c}_k^T (\mathbf{y} - \mathbf{d}_k) \beta_k \right\} - \mathbb{E}_q \left(\frac{\lambda}{\sigma} \right) \left\{ |\beta_k| + \mathbf{1}^T \Psi(\mathbf{a}_k \beta_k + \mathbf{b}_k, \boldsymbol{\sigma}_{-k|k}^2) \right\} \right] \end{aligned}$$

where $\mathbb{E}_q(\sigma^{-1}) = \Gamma(\tilde{A} + 1/2)/\Gamma(\tilde{A})\sqrt{\tilde{B}}$, $\Psi(\mu, \sigma^2) = \mu\{\Phi(\mu/\sigma) - \Phi(-\mu/\sigma)\} + 2\sigma\phi(\mu/\sigma)$ is the expectation of $|Z|$ for $Z \sim N(\mu, \sigma^2)$, and $\boldsymbol{\sigma}_{-k|k}^2 = \text{dg}(\tilde{\boldsymbol{\Sigma}}_{-k|k})$.

After normalizing r_k , we compute $\tilde{\mu}_k = \mathbb{E}_{r_k}(\beta_k)$, and $\tilde{\Sigma}_k = \mathbb{V}_{r_k}(\beta_k)$, and update the Gaussian propagation approximation for $\boldsymbol{\beta}$ using the GMP equations from Section 3.4. The local density r_k can also be retained as the marginal approximation for β_k . For a single coordinate k , the local density can be normalized in $O(np + p^2 + Np)$ time, where N is the number of quadrature nodes (see Supplementary Material C). A full sweep over all p coefficients has cost $O(np^2 + p^3 + Np^2)$.

5 Numerical experience

We do not consider the linear regression or multivariate normal model examples here, since our MP approaches for these models are exact. For the remaining models, we consider MCMC as a gold standard where we use HMC (Hamiltonian Monte Carlo, with the no-U-turn sampler) via `stan` (Hoffman and Gelman, 2014; Stan Development Team, 2025). For each model, we obtained a large number of HMC samples (5×10^5 samples; 5000 warm-up) as the gold standard through the R package `rstan`. Separately, we obtain a short run of HMC samples (5,000 samples; 1,000 warm-up) representing a bare minimum “out of the box” MCMC approach for comparisons with various ABI methods. We use several metrics to assess the quality of marginal posterior density approximations. We use the accuracy metric of Faes et al. (2011) for the j -th marginal posterior density,

$$\text{accuracy}_j = 1 - \frac{1}{2} \int |p(\theta_j|\mathcal{D}) - q(\theta_j)| d\theta_j, \quad j = 1, \dots, p, \quad (18)$$

where $p(\theta_j|\mathcal{D})$ is estimated from a kernel density estimate from the long run of HMC. The quantity accuracy_j takes values between 0 and 1 and is expressed as a percentage. For each dataset and method, we record the minimum (MIN), first quartile (Q1), and median (Q2) of the p accuracies. Accuracies above 95% are usually sufficiently accurate for most univariate inferences. In addition to the accuracy, we record $\sqrt{[\sum_j \{\mathbb{E}_q(\theta_j) - \mathbb{E}(\theta_j|\mathcal{D})\}^2]/p}$ denoted $\text{RMSE}(\mu)$, $\sqrt{[\sum_j \{\text{sd}_q(\theta_j) - \text{sd}(\theta_j|\mathcal{D})\}^2]/p}$ denoted $\text{RMSE}(\sigma)$ for each coefficient, where again posterior quantities are estimated using `stan` using a large number of samples. Lastly, we record $\|\mathbb{V}(\boldsymbol{\theta}|\mathcal{D}) - \mathbb{V}_q(\boldsymbol{\theta})\|_F$ using the Frobenius norm as a multivariate metric.

Where appropriate, we compare our methodology against several ABI methods, including MFVB, Laplace approximation, Expectation Propagation (EP) (Minka, 2001), and Gaussian Variational Bayes (GVB) (Opper and Archambeau, 2009) where the ELBO is approximated via n univariate integrals using numerical quadrature and optimized via the limited memory BFGS method of R’s `optim` function. GMP results initialized from the MFVB, Laplace approximation, GVB, and EP are denoted GMP-M, GMP-L, GMP-G and GMP-E respectively. All methods are implemented using `Rcpp` (Eddelbuettel and François, 2011), and `RcppArmadillo`

(Eddelbuettel, Francois, et al., 2025) to facilitate comparisons. All computations were performed on a computer with a 2.8 GHz Intel Core i7 processor and 32 GB of RAM to perform the computations.

5.1 Multivariate normal model with missing data

For this model we perform a short simulation study to compare our MFMP approximation against unblocked MFVB approximations where $q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X}_{\mathcal{M}}) = q(\boldsymbol{\mu})q(\boldsymbol{\Sigma}) \prod_{i \in \mathcal{M}} q(\mathbf{x}_{i, \mathcal{M}_i})$ and blocked MFVB approximations where $q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X}_{\mathcal{M}}) = q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i \in \mathcal{M}} q(\mathbf{x}_{i, \mathcal{M}_i})$. In the simulation study we conducted 25 replications where $p \in \{4, 8, 16\}$, $n = p \times \{3, 5, 10\}$, the true covariance matrix $\boldsymbol{\Sigma}$ had an AR(1) correlation structure with $\rho \in \{0, 0.5, 0.9\}$, where elements of $\mathbf{x}_i$ were missing completely at random with probabilities $\{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. Marginal posterior accuracies of elements of $\boldsymbol{\mu}$, diagonal elements of $\boldsymbol{\Sigma}$, and off-diagonal elements of $\boldsymbol{\Sigma}$ were recorded separately.

Table 1 summarizes the coefficients based on linear regression fits with ABI method, value of n , value of p , correlation ρ , and missingness percentages as predictors. This suggests that the MFMP method is 5.6% more accurate for the posterior mean parameter μ_j across all simulations, and is roughly 3% more accurate for the posterior for elements of $\boldsymbol{\Sigma}$. Increasing n and/or ρ increases accuracy of all methods on average, increasing p and or the percent of missing values decreases the accuracy of all methods on average. The average times where for the short run Gibbs sampling (0.754s), MFVB (0.004s), MFVB-blocked (0.012s) and MFMP (0.0095s). Thus, MFMP represents a good accuracy versus speed trade-off.

5.2 Generalized Linear Models

In this section, we compare several different approaches to fitting the Bayesian binary regression models via probit and logistic regression on commonly used benchmark datasets **Breastfeed** ($n = 135, p = 10$), **Ionosphere**, ($n = 351, p = 33$) **Spam** ($n = 4601, p = 58$), and **Caravan** ($n = 5822, p = 86$). The dataset **Breastfeed** is based on the study of Moustaki et al. (1998) and is available from the website of the book of Heritier et al. (2009). The datasets **Ionosphere** and **Spam** can be found in the UCI Machine Learning Repository (Dua and

Coefficient	μ_j	Diagonal Σ	Off-diagonal Σ
MFVB	93.85 (0.23)	97.79 (0.29)	95.59 (0.31)
MFVB-blocked	+3.82 (0.16)	+0.69 (0.20)	+0.63 (0.21)
MFMP	+5.64 (0.16)	+2.89 (0.20)	+3.00 (0.22)
Gibbs	+5.76 (0.16)	+4.14 (0.22)	+4.97 (0.25)
n	+0.02 (0.00)	+0.02 (0.00)	+0.01 (0.00)
p	-0.11 (0.02)	-0.23 (0.02)	-0.03 (0.02)
ρ	+2.86 (0.18)	+4.43 (0.23)	+6.64 (0.24)
Miss. %	-16.08 (0.77)	-26.04 (0.97)	-32.14 (0.10)

Table 1: Summary of coefficients based on linear regression fits using accuracy as the response across all simulation settings grouped by posterior mean μ_i , posterior of diagonal elements of Σ , and off-diagonal elements of Σ . First row represents baseline accuracy, subsequent rows indicate adjustments from baseline accuracy. Bracketed values represent standard errors.

Graff, 2017). The dataset **Caravan** can be found from the **ISLR** package in R (James et al., 2021).

Table 2 summarize the performance metrics described in Section 5 for probit models. Additional results for logistic regression models can be found in Supplementary Material H. However, these results are similar to those for the probit model case. The MP entry in Table 2 corresponds to the method outlined in Section 4.4 and generally performs similarly to other Gaussian-based ABI methods. For probit regression, MFVB is based on Consonni and Marin (2007). For almost all models and datasets, either EP or one of the GMP variants are the most accurate with competitive RMSE’s and Frob. metrics while being tens of times faster than a short-run of **stan**. Most GMP-based approximations improve significantly over the respective method they were initialized with. Occasionally, our GMP method performed poorly due to numerical instability (see GMP-G for probit regression for the **Caravan** dataset). GMP is comparable in speed to other ABI methods.

5.3 Bayesian Lasso regression

Here, we compare MFVB and GVB (described in Supplementary Material I) with GMP initialized with these methods. The parameter λ is chosen using the Gibbs sampler described in Park and Casella (2008) using the prior $\lambda^2 \sim \text{Gamma}(10^{-4}, 10^{-4})$ and taking the posterior mean of λ based on 5000 samples after a burn-in of 1000 samples. Table 3 summarizes the metrics described in Section 5 using the **Hitters** ($n = 263$, $p = 19$) (ISLR package; (James et al.,

Dataset	Method	RMSE(μ)	RMSE(σ)	MIN	Q1.	Q2.	Frob.	Time (s)
Breastfeed	stan	0.0093	0.0044	97.6	97.8	98.2	0.0031	0.768
	MFVB	0.0846	0.1910	65.3	68.0	73.5	0.0615	0.000
	Laplace	0.0846	0.0124	88.9	92.8	95.1	0.0056	0.000
	EP	0.0011	0.0060	96.4	98.3	99.2	0.0030	0.459
	GVB	0.0012	0.0073	96.4	98.1	99.1	0.0036	0.012
	MP	0.0012	0.0037	96.5	98.3	99.2	0.0018	0.000
	GMP-L	0.0155	0.0079	98.9	99.0	99.1	0.0035	0.004
	GMP-G	0.0011	0.0038	99.3	99.5	99.6	0.0023	0.002
	GMP-E	0.0009	0.0038	99.3	99.5	99.6	0.0021	0.001
Ion.	stan	0.004	0.003	97.4	97.9	98.2	0.0012	2.78
	MFVB	0.114	0.171	33.5	48.5	52.4	0.0170	0.02
	Laplace	0.114	0.011	60.4	79.1	89.1	0.0023	0.00
	EP	0.002	0.004	97.0	98.3	98.9	0.0010	0.55
	GVB	0.003	0.013	95.7	96.9	97.3	0.0021	0.12
	MP	0.003	0.009	96.5	97.6	98.0	0.0017	0.00
	GMP-L	0.015	0.019	94.3	95.6	95.9	0.0034	0.06
	GMP-G	0.003	0.013	96.7	97.1	97.6	0.0020	0.03
	GMP-E	0.004	0.014	95.9	96.9	97.3	0.0023	0.02
Spam	stan	0.0019	0.0026	96.6	97.8	98.0	0.0004	40.1
	MFVB	0.1180	0.0948	2.7	40.2	52.8	0.0051	0.6
	Laplace	0.0319	0.0065	75.8	93.1	96.4	0.0004	0.2
	EP	0.0013	0.0056	82.4	96.8	98.4	0.0004	2.0
	GVB	0.0031	0.0097	80.3	96.1	98.4	0.0007	2.5
	MP	0.0031	0.0093	80.8	96.1	98.4	0.0007	0.1
	GMP-L	0.0050	0.0047	96.5	98.6	98.9	0.0005	1.1
	GMP-G	0.0023	0.0033	97.2	99.1	99.2	0.0005	0.6
	GMP-E	0.0016	0.0036	96.8	99.0	99.2	0.0003	0.3
Caravan	stan	0.018	0.025	95.2	97.2	97.7	0.023	1653.5
	MFVB	0.412	0.604	13.7	55.5	60.7	0.192	1.1
	Laplace	0.407	0.109	53.2	94.4	96.7	0.044	0.6
	EP	0.025	0.074	81.1	96.9	98.0	0.034	3.5
	GVB	0.267	0.396	21.2	27.7	40.2	0.119	6.4
	MP	0.455	0.361	18.7	55.8	65.5	0.156	3.5
	GMP-L	0.098	0.111	85.4	96.1	97.7	0.058	8.7
	GMP-G	1.100	0.521	0.5	31.2	44.2	0.161	21.4
	GMP-E	0.070	0.098	86.9	97.5	98.1	0.053	7.0

Table 2: Performance metrics for Probit regression.

2021)), **Crime** ($n = 2215, p = 98$) (UCI Machine Learning Repository; (Redmond, 2002)) and **Eye** ($n = 120, p = 200$) (flare package; (Li et al., 2022)) datasets.

Table 3 shows that for each dataset almost all marginal posterior density approximations for both GMP based approximations are nearly visually indistinguishable from those based on **stan** whilst being at least tens of times faster than the corresponding short run MCMC approach.

Data	Method	RMSE(μ)	RMSE(σ)	MIN	Q1.	Q2.	Frob.	Time (s)
Hitters	stan	0.0926	0.0177	87.5	94.0	96.1	0.0342	28.94
	MFVB	0.0913	0.0144	31.7	69.8	85.6	0.0197	0.01
	GVB	0.0019	0.0075	86.3	89.9	92.1	0.0099	0.01
	GMP-M	0.0024	0.0042	96.8	97.8	98.2	0.0052	0.03
	GMP-G	0.0093	0.0030	97.7	98.6	99.2	0.0038	0.01
Crime	stan	0.0939	0.0041	43.9	83.3	89.4	0.0034	93.60
	MFVB	0.0140	0.0033	36.1	82.0	88.4	0.0038	0.03
	GVB	0.0002	0.0013	79.1	94.0	95.6	0.0019	0.03
	GMP-M	0.0005	0.0007	95.1	98.8	99.0	0.0011	0.87
	GMP-G	0.0010	0.0007	95.9	99.6	99.7	0.0008	0.60
Eye	stan	0.0183	0.0022	94.8	96.8	97.3	0.0090	34.34
	MFVB	0.0138	0.0056	51.5	79.8	83.3	0.0085	0.33
	GVB	0.0013	0.0036	83.3	87.7	88.7	0.0048	0.08
	GMP-M	0.0012	0.0058	92.7	93.5	93.7	0.0064	1.95
	GMP-G	0.0069	0.0027	99.1	99.4	99.5	0.0042	3.16

Table 3: Performance metrics for Bayesian Lasso.

6 Conclusion

We have introduced moment propagation (MP), a framework for improving marginal posterior approximations by propagating conditional posterior moment information between parameter blocks. In conjugate settings, MP identifies variance terms omitted by MFVB and uses moment matching to construct corrected marginal updates. We developed both mean-field and Gaussian variants, showed exact marginal posterior recovery for linear regression and multivariate normal models, and established asymptotically correct posterior moments in two-component settings. Across the examples considered, MP-based methods, particularly GMP, often produced substantially more accurate marginal posterior approximations than MFVB and other approximate Bayesian inference methods while remaining much faster than MCMC.

Several challenges remain. GMP can depend on the quality of the initial Gaussian approximation, and numerical integration may introduce additional cost or instability in some models. Theoretical properties are currently best understood in conjugate two-component settings, leaving open questions about convergence, rates, and accuracy in more general multi-component and non-conjugate models. Future work will focus on these theoretical issues, on improved initialization, stability, and acceleration strategies, and on extensions to richer hierarchical models, including random-effects and state-space settings.

7 Acknowledgments

The following sources of funding are gratefully acknowledged: Australian Research Council Discovery Project grant (DP210100521) to JO. We also thank Prof. Matt P. Wand for making comments on an early draft of the paper.

References

- Albert, J. H. and S. Chib (1993). “Bayesian analysis of binary and polychotomous response data”. In: *Journal of the American Statistical Association* 88, pp. 669–679.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.
- Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). “Variational inference: A review for statisticians”. In: *Journal of the American Statistical Association* 112.518, pp. 859–877.
- Challis, E. and D. Barber (2013). “Gaussian Kullback-Leibler approximate inference”. In: *Journal of Machine Learning Research* 14.32, pp. 2239–2286.
- Consonni, G. and J.-M. Marin (2007). “Mean-field variational approximate Bayesian inference for latent variable models”. In: *Computational Statistics & Data Analysis* 52.2, pp. 790–798. ISSN: 0167-9473.
- Dua, D. and C. Graff (2017). *UCI Machine Learning Repository*.
- Eddelbuettel, D., R. Francois, et al. (2025). *RcppArmadillo: 'Rcpp' Integration for the 'Armadillo' Templated Linear Algebra Library*. R package version 14.6.0-1.
- Eddelbuettel, D. and R. François (2011). “Rcpp: Seamless R and C++ Integration”. In: *Journal of Statistical Software* 40.8, pp. 1–18.
- Faes, C., J. Ormerod, and M. Wand (Sept. 2011). “Variational Bayesian inference for parametric and nonparametric regression With missing data”. In: *Journal of the American Statistical Association* 106, pp. 959–971.
- Giordano, R., T. Broderick, and M. I. Jordan (2015). “Linear response methods for accurate covariance estimates from mean field variational Bayes”. In: *NIPS*, pp. 1441–1449.
- Heritier, S. et al. (2009). *Robust methods in biostatistics*. John Wiley & Sons.
- Hoffman, M. D. and A. Gelman (2014). “The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo”. In: *Journal of Machine Learning Research* 15.47, pp. 1593–1623.
- James, G. et al. (2021). *ISLR: Data for an Introduction to Statistical Learning with Applications in R*. R package version 1.4.

- Kullback, S. and R. A. Leibler (1951). “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22, pp. 79–86.
- Li, X. et al. (2022). *flare: Family of Lasso Regression*. R package version 1.7.0.1.
- Minka, T. P. (2001). “Expectation propagation for approximate Bayesian inference”. In: UAI’01. Seattle, Washington, pp. 362–369.
- (2005). *Divergence Measures and Message Passing*. Technical Report MSR-TR-2005-173. Microsoft Research, pp. 1–17.
- Moustaki, I., M. Victoria-Feser, and H. Hyams (1998). “A UK study on the effect of socioeconomic background of pregnant women and hospital practice on the decision to breastfeed and the initiation and duration of breastfeeding”. In: *Statistics Research Report LSERR44, London School of Economics, London WC2A 2AE*.
- Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press.
- Opper, M. and C. Archambeau (2009). “The variational Gaussian approximation revisited”. In: *Neural Computation* 21.3, pp. 786–792.
- Ormerod, J. T. and M. P. Wand (2010). “Explaining variational approximations”. In: *The American Statistician* 64, pp. 140–153.
- Park, T. and G. Casella (2008). “The Bayesian Lasso”. In: *Journal of the American Statistical Association* 103.482, pp. 681–686.
- Redmond, M. (2002). *Communities and Crime*. UCI Machine Learning Repository.
- Rohde, D. and M. P. Wand (2016). “Semiparametric Mean Field Variational Bayes: General Principles and Numerical Issues”. In: *Journal of Machine Learning Research* 17.172, pp. 1–47.
- Rue, H., S. Martino, and N. Chopin (2009). “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, pp. 319–392.
- Stan Development Team (2025). *RStan: the R interface to Stan*. R package version 2.32.7.
- Turner, R. et al. (Jan. 2011). “Two problems with variational expectation maximisation for time-series models”. In: *Bayesian Time Series Models*, pp. 109–130.
- Wang, B. and D. M. Titterton (Jan. 2005). “Inadequacy of interval estimates corresponding to variational Bayesian approximations”. In: *AISTATS 2005 - Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*.
- Wang, C. and D. M. Blei (2013). “Variational inference in nonconjugate models”. In: *Journal of Machine Learning Research* 14.1, pp. 1005–1031. ISSN: 1532-4435.

Supplementary material to “Moment Propagation”

by John T. Ormerod, Weichang Yu, Mohammad Javad Davoudabadi, & Yuhao Li

This material will appear as an online supplement.

A Moment Results

This appendix collects moment identities used in the derivations of the MP updates. Sections [A.1](#) and [A.2](#) are used for quadratic forms appearing in the linear regression and multivariate normal examples. Section [A.3](#) gives inverse-Wishart moment identities used in the multivariate normal examples. Section [A.4](#) records block inverse-Wishart identities needed for the missing-data updates.

Throughout this appendix, $\text{dg}(\mathbf{A})$ denotes the vector of diagonal elements, $\mathbb{E}\mathbb{W}\mathbb{V}(\mathbf{A})$ denotes the matrix of element-wise variances, $\odot$ denotes the Hadamard product, and $\mathbf{K}_{p,p}$ denotes the commutation matrix.

A.1 Moments for the Multivariate Gaussian Distribution

It is shown in Mathai and Provost (1992) that the s -th cumulant of $\mathbf{x}^T \mathbf{A} \mathbf{x}$ where $\mathbf{A}$ is a $p \times p$ symmetric matrix and $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by

$$\kappa(s) = 2^{s-1} s! \left[\frac{\text{tr}\{(\mathbf{A}\boldsymbol{\Sigma})^s\}}{s} + \boldsymbol{\mu}^T (\mathbf{A}\boldsymbol{\Sigma})^{s-1} \mathbf{A} \boldsymbol{\mu} \right].$$

The h -th moment of $\mathbf{x}^T \mathbf{A} \mathbf{x}$ can be calculated recursively using these cumulants via

$$\mathbb{E}[(\mathbf{x}^T \mathbf{A} \mathbf{x})^h] = \sum_{i=0}^{h-1} \frac{(h-1)!}{(h-1-i)!i!} \kappa(h-i) \mathbb{E}[(\mathbf{x}^T \mathbf{A} \mathbf{x})^i]$$

with $\mathbb{E}[(\mathbf{x}^T \mathbf{A} \mathbf{x})^0] = 1$. We can use the above moments to derive

$$\begin{aligned}\mathbb{E}[(\mathbf{x}^T \mathbf{A} \mathbf{x})^2] &= 2\text{tr}(\mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\Sigma}) + 4\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\mu} + [\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} + \text{tr}(\mathbf{A} \boldsymbol{\Sigma})]^2, \\ \mathbb{V}(\mathbf{x}^T \mathbf{A} \mathbf{x}) &= 2\text{tr}(\mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\Sigma}) + 4\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\mu}, \quad \text{and} \\ \text{Cov}(\mathbf{x}^T \mathbf{A} \mathbf{x}, \mathbf{x}^T \mathbf{B} \mathbf{x}) &= 2\text{tr}(\mathbf{A} \boldsymbol{\Sigma} \mathbf{B} \boldsymbol{\Sigma}) + 4\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\Sigma} \mathbf{B} \boldsymbol{\mu}.\end{aligned}\tag{19}$$

A.2 Moments for the Multivariate t Distribution

The p -variate t -distribution, denoted as $\mathbf{x} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, has the density function given by

$$p(\mathbf{x}) = \frac{\Gamma[(\nu + p)/2]}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}\det(\boldsymbol{\Sigma})^{1/2}} \left[1 + \frac{1}{\nu}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-(\nu+p)/2}$$

where $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and ν are the location, scale and degrees of freedom parameters respectively. The mean and variance of $\mathbf{x}$ are given by $\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}$ (provided $\nu > 1$), covariance $\mathbb{V}(\mathbf{x}) = \nu \boldsymbol{\Sigma} / (\nu - 2)$ (provided $\nu > 2$). Fourth moments require $\nu > 4$.

Theorem 2.1 of Rong et al. (2012) If $\mathbf{x} \sim t_p(\boldsymbol{\mu}, a\boldsymbol{\Sigma}, b)$ and $\mathbf{A}$ is a symmetric matrix then

$$\begin{aligned}\mathbb{E}[(\mathbf{x}^T \mathbf{A} \mathbf{x})^2] &= \frac{a^2 b^2}{(b-2)(b-4)} [2\text{tr}(\mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\Sigma}) + \text{tr}(\mathbf{A} \boldsymbol{\Sigma})^2] \\ &\quad + \frac{4ab}{b-2} \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\mu} + (\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu})^2 + \frac{2ab}{b-2} (\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu}) \text{tr}(\mathbf{A} \boldsymbol{\Sigma}),\end{aligned}\tag{20}$$

$$\mathbb{V}(\mathbf{x}^T \mathbf{A} \mathbf{x}) = \frac{2a^2 b^2 \text{tr}(\mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\Sigma})}{(b-2)(b-4)} + \frac{2a^2 b^2 [\text{tr}(\mathbf{A} \boldsymbol{\Sigma})]^2}{(b-2)^2(b-4)} + \frac{4ab}{b-2} \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\mu},\tag{21}$$

and

$$\text{Cov}(\mathbf{x}^T \mathbf{A} \mathbf{x}, \mathbf{x}^T \mathbf{B} \mathbf{x}) = \frac{2a^2 b^2 \text{tr}(\mathbf{A} \boldsymbol{\Sigma} \mathbf{B} \boldsymbol{\Sigma})}{(b-2)(b-4)} + \frac{2a^2 b^2 \text{tr}(\mathbf{A} \boldsymbol{\Sigma}) \text{tr}(\mathbf{B} \boldsymbol{\Sigma})}{(b-2)^2(b-4)} + \frac{4ab(\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\Sigma} \mathbf{B} \boldsymbol{\mu})}{b-2}\tag{22}$$

provided $b > 4$.

A.3 Moments for the Inverse-Wishart distribution

We use the parameterization of the inverse-Wishart distribution to have the following density

$$p(\mathbf{\Sigma}) = \frac{\det(\mathbf{\Psi})^{d/2}}{2^{dp/2}\Gamma_p(d/2)} |\mathbf{\Sigma}|^{-(d+p+1)/2} \exp\left[-\frac{1}{2}\text{tr}(\mathbf{\Psi}\mathbf{\Sigma}^{-1})\right] \cdot \mathbb{I}(\mathbf{\Sigma} \in \mathbb{S}_p^+), \quad (23)$$

where

$$\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma\{a + (1-j)/2\}$$

is the multivariate gamma function.. The following Theorem and Corollary can be found in von Rosen (1988).

Theorem 3.1 of von Rosen (1988): Let $\mathbf{\Sigma} \sim \text{IW}_p(\mathbf{\Psi}, d)$ then

$$\text{i. } \mathbb{E}(\mathbf{\Sigma}) = \frac{\mathbf{\Psi}}{d-p-1}, \quad \text{if } d-p-1 > 0;$$

Corollary 3.1 of von Rosen (1988): Let $\mathbf{\Sigma} \sim \text{IW}_p(\mathbf{\Psi}, d)$ with

$$\begin{aligned} c_2^{-1} &= (d-p)(d-p-1)(d-p-3), & c_1 &= (d-p-2)c_2, \\ c_3 &= \frac{d-p-3}{(d-p-5)(d-p+1)}, & \text{and} & \quad c_4 = \frac{2}{(d-p-5)(d+p+1)}. \end{aligned}$$

Then,

(i) if $d-p-3 > 0$,

$$\mathbb{E}[\mathbf{\Sigma}^2] = (c_1 + c_2)\mathbf{\Psi}^2 + c_2\mathbf{\Psi}\text{tr}(\mathbf{\Psi});$$

(ii) if $d-p-5 > 0$,

$$\begin{aligned} \mathbb{E}[\mathbf{\Sigma}^3] &= (c_3c_1 + c_3c_2 + c_4c_1 + 5c_4c_2)\mathbf{\Psi}^3 + (2c_3c_2 + c_4c_1 + c_4c_2)\text{tr}(\mathbf{\Psi})\mathbf{\Psi}^2 \\ &\quad - (c_3c_2 + c_4c_2)\text{tr}(\mathbf{\Psi}^{-2})\mathbf{\Psi} - c_4c_2(\text{tr}(\mathbf{\Psi}))^2\mathbf{\Psi}; \end{aligned}$$

(iii) if $d-p-5 > 0$,

$$\begin{aligned} \mathbb{E}[\text{tr}(\mathbf{\Sigma}^2)\mathbf{\Sigma}] &= c_4(c_1 + c_2)\mathbf{\Psi}^3 + (c_4c_1 + c_3c_2)[\text{tr}(\mathbf{\Psi})]^2\mathbf{\Psi} \\ &\quad + (c_3c_1 + c_3c_2 + c_4c_2)\text{tr}(\mathbf{\Psi}^2)\mathbf{\Psi} - 2c_4c_2\text{tr}(\mathbf{\Psi})\mathbf{\Psi}^2; \end{aligned}$$

(iv) if $d - p - 5 > 0$,

$$\mathbb{E} [\text{tr}(\mathbf{\Sigma})^3] = c_3 c_1 [\text{tr}(\mathbf{\Psi})^3] + 4c_4 c_2 \text{tr}(\mathbf{\Psi}^3) + 2(c_4 c_1 + c_3 c_2) \text{tr}(\mathbf{\Psi}^2) \text{tr}(\mathbf{\Psi});$$

(v) if $d - p - 3 > 0$,

$$\mathbb{E} [\text{tr}(\mathbf{\Sigma}) \mathbf{\Sigma}] = c_1 \text{tr}(\mathbf{\Psi}) \mathbf{\Psi} + 2c_2 \mathbf{\Psi}^2$$

(vi) if $d - p - 1 > 0$,

$$\mathbb{E} [\mathbf{\Sigma} \otimes \mathbf{\Sigma}^{-1}] = \frac{d}{d - p - 1} \mathbf{\Psi} \otimes \mathbf{\Psi}^{-1} - \frac{1}{d - p - 1} [\text{vec}(\mathbf{I}_p) \text{vec}(\mathbf{I}_p)^T + \mathbf{K}_{p,p}]; \quad \text{and}$$

(vii) if $d - p - 1 > 0$,

$$\mathbb{E} [\text{tr}(\mathbf{\Sigma}) \mathbf{\Sigma}^{-1}] = \frac{d}{d - p - 1} \text{tr}(\mathbf{\Psi}) \mathbf{\Psi}^{-1} - \frac{2}{d - p - 1} \mathbf{I}_p.$$

In von Rosen (1988), the coefficient multiplying $\mathbf{\Psi}^2$ in part (i) appears to contain a typographical error. The expression displayed above is the version used in our derivations. The variances and covariances for elements of $\mathbf{\Sigma}$ are given by

$$\begin{aligned} \mathbb{V}(\Sigma_{ij}) &= \frac{(d - p + 1)\psi_{ij}^2 + (d - p - 1)\psi_{ii}\psi_{jj}}{(d - p)(d - p - 1)^2(d - p - 3)}, \\ \mathbb{V}(\Sigma_{ii}) &= \frac{2\psi_{ii}^2}{(d - p - 1)^2(d - p - 3)}, \quad \text{and} \\ \text{Cov}(\Sigma_{ij}, \Sigma_{kl}) &= \frac{2\psi_{ij}\psi_{kl} + (d - p - 1)(\psi_{ik}\psi_{jl} + \psi_{il}\psi_{kj})}{(d - p)(d - p - 1)^2(d - p - 3)}. \end{aligned} \tag{24}$$

provided $d - p - 3 > 0$ (see Press, 2005). Note that the element-wise variance matrix for $\mathbf{\Sigma}$ is

$$\mathbb{E} \mathbb{W} \mathbb{V}(\mathbf{\Sigma}) = \frac{(d - p + 1)(\mathbf{\Psi} \odot \mathbf{\Psi})}{(d - p)(d - p - 1)^2(d - p - 3)} + \frac{(d - p - 1) \text{dg}(\mathbf{\Psi}) \text{dg}(\mathbf{\Psi})^T}{(d - p)(d - p - 1)^2(d - p - 3)}$$

whose diagonal elements are

$$\text{dg}[\mathbb{E} \mathbb{W} \mathbb{V}(\mathbf{\Sigma})] = \frac{2 \text{dg}(\mathbf{\Psi})^{\odot 2}}{(d - p - 1)^2(d - p - 3)}.$$

A.4 Block Inverse-Wishart identities

In Section 4.3 some identities concerning block inverse-Wishart distributions are needed in the derivation $\mathbb{E}_r(\mathbf{x}_{i,M_i})$ and $\mathbb{V}_r(\mathbf{x}_{i,M_i})$ needed to update $q(\mathbf{x}_{i,M_i})$. The update for $q(\mathbf{x}_{i,M_i})$ requires expectations involving the Gaussian conditional regression matrix $\Sigma_{M_i O_i} \Sigma_{O_i O_i}^{-1}$ and the Schur complement $\Sigma_{M_i|O_i} = \Sigma_{M_i M_i} - \Sigma_{M_i O_i} \Sigma_{O_i O_i}^{-1} \Sigma_{O_i M_i}$. The following proposition collects the Inverse-Wishart identities used in Section 4.3 and Section F.

Proposition. Let $\Sigma \sim IW_p(\Psi, d)$, and let M and O be non-empty disjoint index sets forming a partition of $\{1, \dots, p\}$, with $m = |M|$ and $o = |O|$. Define $\Sigma_{M|O} = \Sigma_{MM} - \Sigma_{MO} \Sigma_{OO}^{-1} \Sigma_{OM}$, and $\Psi_{M|O} = \Psi_{MM} - \Psi_{MO} \Psi_{OO}^{-1} \Psi_{OM}$. The following proposition assumes $O \neq \emptyset$. The case $O = \emptyset$ is handled separately in the main text, where $\Sigma_{M|O}$ reduces to Σ_{MM} . Then,

$$\Sigma_{M|O} \sim IW_m(\Psi_{M|O}, d), \quad \text{and} \quad \Sigma_{MO} \Sigma_{OO}^{-1} \mid \Sigma_{M|O} \sim MN_{m \times o}(\Psi_{MO} \Psi_{OO}^{-1}, \Sigma_{M|O}, \Psi_{OO}^{-1})$$

where $MN_{m \times o}(\mathbf{A}, \mathbf{B}, \mathbf{C})$ denotes the matrix-variate normal distribution. Consequently,

$$\mathbb{E}(\Sigma_{MO} \Sigma_{OO}^{-1}) = \Psi_{MO} \Psi_{OO}^{-1}, \quad \text{and} \quad \mathbb{E}(\Sigma_{M|O}) = \frac{\Psi_{M|O}}{d - m - 1}$$

provided $d > m + 1$. Moreover, for any fixed o -vector $\mathbf{x}$,

$$\mathbb{V}(\Sigma_{MO} \Sigma_{OO}^{-1} \mathbf{x}) = \frac{\mathbf{x}^T \Psi_{OO}^{-1} \mathbf{x}}{d - m - 1} \Psi_{M|O}.$$

Proof sketch. The distributional statements follow from Theorem 3 of Bodnar and Okhrin (2008), after relabelling the blocks. The first two moment identities follow directly. For the final identity, condition on $\Sigma_{M|O}$:

$$\Sigma_{MO} \Sigma_{OO}^{-1} \mathbf{x} \mid \Sigma_{M|O} \sim N_m(\Psi_{MO} \Psi_{OO}^{-1} \mathbf{x}, (\mathbf{x}^T \Psi_{OO}^{-1} \mathbf{x}) \Sigma_{M|O}).$$

The conditional mean is non-random, so the law of total variance gives

$$\mathbb{V}(\Sigma_{MO} \Sigma_{OO}^{-1} \mathbf{x}) = (\mathbf{x}^T \Psi_{OO}^{-1} \mathbf{x}) \mathbb{E}(\Sigma_{M|O}) = \frac{\mathbf{x}^T \Psi_{OO}^{-1} \mathbf{x}}{d - m - 1} \Psi_{M|O}.$$

B Proof of Theorems 1 to 3

Recap of MFMP setup

We approximate each marginal posterior of each subvector $p(\boldsymbol{\theta}_k \mid \mathcal{D})$ with $q_k(\boldsymbol{\theta}_k; \hat{\boldsymbol{\lambda}}_k) \in \mathcal{Q}_k = \{q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k) : \boldsymbol{\lambda}_k \in \boldsymbol{\Lambda}_k\}$, where each $\mathcal{Q}_k$ is a convenient parametric family of distributions. Here, we specify $\{\hat{\boldsymbol{\lambda}}_k\}_{k=1}^K$ to solve the system of integral equations for $k = 1, \dots, K$:

$$\begin{aligned} \int \boldsymbol{\theta}_k q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k) d\boldsymbol{\theta}_k &= \int \boldsymbol{\theta}_k \int^{\otimes^{K-1}} p(\boldsymbol{\theta}_k \mid \mathcal{D}, \boldsymbol{\theta}_{-k}) \prod_{j \neq k} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j d\boldsymbol{\theta}_k \\ \int \boldsymbol{\theta}_k \boldsymbol{\theta}_k^\top q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k) d\boldsymbol{\theta}_k &= \int \boldsymbol{\theta}_k \boldsymbol{\theta}_k^\top \int^{\otimes^{K-1}} p(\boldsymbol{\theta}_k \mid \mathcal{D}, \boldsymbol{\theta}_{-k}) \prod_{j \neq k} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j d\boldsymbol{\theta}_k \\ \int \mathbf{s}_k(\boldsymbol{\theta}_k) q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k) d\boldsymbol{\theta}_k &= \int \mathbf{s}_k(\boldsymbol{\theta}_k) \int^{\otimes^{K-1}} p(\boldsymbol{\theta}_k \mid \mathcal{D}, \boldsymbol{\theta}_{-k}) \prod_{j \neq k} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j d\boldsymbol{\theta}_k \end{aligned}$$

Essentially, we are choosing the q_k 's such that their respective moments $\mathbb{E}(\boldsymbol{\theta}_k)$, $\mathbb{E}(\boldsymbol{\theta}_k \boldsymbol{\theta}_k^\top)$, and $\mathbb{E}[\mathbf{s}_k(\boldsymbol{\theta}_k)]$ match those based on $\tilde{r}_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_{-k}) = \int^{\otimes^{K-1}} p(\boldsymbol{\theta}_k \mid \mathcal{D}, \boldsymbol{\theta}_{-k}) \prod_{j \neq k} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j$. We mandate the matching of the first and second order moment of each $\boldsymbol{\theta}_k$, while allowing the users flexibility to match other suitable moments through $\mathbf{s}_k$. If no additional moments are required to be matched, then $\mathbf{s}_k(\boldsymbol{\theta}_k) = 1$. To obtain the solution $\{\hat{\boldsymbol{\lambda}}_k\}_{k=1}^K$, we iterate through a fixed point algorithm. More specifically, at iteration $\tau = 1, 2, \dots$, we update $\lambda_k^{(\tau)}$ for $k = 1, \dots, K$ such that

$$\begin{aligned} \int \boldsymbol{\theta}_k q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^{(\tau)}) d\boldsymbol{\theta}_k &= \int \boldsymbol{\theta}_k \int^{\otimes^{K-1}} p(\boldsymbol{\theta}_k \mid \mathcal{D}, \boldsymbol{\theta}_{-k}) \prod_{j \neq k} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j^{(\tau-1)}) d\boldsymbol{\theta}_j d\boldsymbol{\theta}_k \\ \int \boldsymbol{\theta}_k \boldsymbol{\theta}_k^\top q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^{(\tau)}) d\boldsymbol{\theta}_k &= \int \boldsymbol{\theta}_k \boldsymbol{\theta}_k^\top \int^{\otimes^{K-1}} p(\boldsymbol{\theta}_k \mid \mathcal{D}, \boldsymbol{\theta}_{-k}) \prod_{j \neq k} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j^{(\tau-1)}) d\boldsymbol{\theta}_j d\boldsymbol{\theta}_k \\ \int \mathbf{s}_k(\boldsymbol{\theta}_k) q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^{(\tau)}) d\boldsymbol{\theta}_k &= \int \mathbf{s}_k(\boldsymbol{\theta}_k) \int^{\otimes^{K-1}} p(\boldsymbol{\theta}_k \mid \mathcal{D}, \boldsymbol{\theta}_{-k}) \prod_{j \neq k} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j^{(\tau-1)}) d\boldsymbol{\theta}_j d\boldsymbol{\theta}_k \end{aligned}$$

Auxiliary notations and assumptions

We define notations:

$$\mathbf{T}_k(\boldsymbol{\theta}_k) = \begin{pmatrix} \boldsymbol{\theta}_k \\ \text{vech}(\boldsymbol{\theta}_k \boldsymbol{\theta}_k^\top) \\ \mathbf{s}_k(\boldsymbol{\theta}_k) \end{pmatrix}$$

and

$$\boldsymbol{\Psi}_n(\boldsymbol{\lambda}) = \mathbf{m}(\boldsymbol{\lambda}) - \mathbf{G}_n(\boldsymbol{\lambda}), \quad \boldsymbol{\Psi}(\boldsymbol{\lambda}) = \mathbf{m}(\boldsymbol{\lambda}) - \mathbf{G}(\boldsymbol{\lambda}),$$

where

$$\mathbf{m}(\boldsymbol{\lambda}) = \begin{pmatrix} \int \mathbf{T}_1(\boldsymbol{\theta}_1) q_1(\boldsymbol{\theta}_1; \boldsymbol{\lambda}_1) d\boldsymbol{\theta}_1 \\ \vdots \\ \int \mathbf{T}_K(\boldsymbol{\theta}_K) q_K(\boldsymbol{\theta}_K; \boldsymbol{\lambda}_K) d\boldsymbol{\theta}_K \end{pmatrix},$$

and

$$\mathbf{G}_n(\boldsymbol{\lambda}) = \begin{pmatrix} \int \mathbf{T}_1(\boldsymbol{\theta}_1) \int^{\otimes^{K-1}} p(\boldsymbol{\theta}_1 \mid \mathcal{D}, \boldsymbol{\theta}_{-1}) \prod_{j \neq 1} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j d\boldsymbol{\theta}_1 \\ \vdots \\ \int \mathbf{T}_K(\boldsymbol{\theta}_K) \int^{\otimes^{K-1}} p(\boldsymbol{\theta}_K \mid \mathcal{D}, \boldsymbol{\theta}_{-K}) \prod_{j \neq K} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j d\boldsymbol{\theta}_K \end{pmatrix}.$$

Denote the solution set by $\mathcal{S}_n = \{\boldsymbol{\lambda}_{1:K} : \boldsymbol{\Psi}_n(\boldsymbol{\lambda}_{1:K}) = \mathbf{0}\}$. Also, we denote the full conditional expectations as

$$\mathbf{h}_{kn}(\boldsymbol{\theta}_{-k}) = \mathbb{E}\{\mathbf{T}_k(\boldsymbol{\theta}_k) \mid \mathcal{D}_n, \boldsymbol{\theta}_{-k}\}.$$

Denote the true value of the parameter subvectors by $\boldsymbol{\theta}_{1:K}^*$. Denote the induced expectation and scaled variance of any point $\boldsymbol{\lambda}_k \in \boldsymbol{\Lambda}_k$ by $\boldsymbol{\mu}_k(\boldsymbol{\lambda}_k) = \int \boldsymbol{\theta}_k q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k) d\boldsymbol{\theta}_k$, $\boldsymbol{\Gamma}_k^{(n)}(\boldsymbol{\lambda}_k) = n \int \{\boldsymbol{\theta}_k - \boldsymbol{\mu}_k(\boldsymbol{\lambda}_k)\} \{\boldsymbol{\theta}_k - \boldsymbol{\mu}_k(\boldsymbol{\lambda}_k)\}^\top q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k) d\boldsymbol{\theta}_k$, and $\mathbf{U}_k(\boldsymbol{\lambda}_k) = \frac{1}{n} \boldsymbol{\Gamma}_k^{(n)}(\boldsymbol{\lambda}_k) + \boldsymbol{\mu}_k(\boldsymbol{\lambda}_k) \boldsymbol{\mu}_k(\boldsymbol{\lambda}_k)^\top$. The theorems presented in this section may require only a subset of the following list of assumptions. The required assumptions for each theorem is stated in their respective technical statements.

(A1) The variational parameter space $\boldsymbol{\Lambda}_{1:K} = \boldsymbol{\Lambda}_1 \times \dots \times \boldsymbol{\Lambda}_K$ is convex and compact.

(A2) There exists a unique $\boldsymbol{\theta}_{1:K}^* = (\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^*) \in \boldsymbol{\Theta}$ and sequence $\epsilon_n \rightarrow 0$ such that

$$p(\|\boldsymbol{\theta}_{1:K} - \boldsymbol{\theta}_{1:K}^*\| > \epsilon_n \mid \mathcal{D}_n) \xrightarrow{\mathbb{P}^*} 0.$$

Moreover, there exists functions $\{\mathbf{h}_k^*(\cdot)\}_{k=1}^K$ such that $\mathbf{h}_k^*(\boldsymbol{\theta}_{-k}^*) = \mathbf{T}_k(\boldsymbol{\theta}_k^*)$ and $\sup_{\boldsymbol{\theta}_{-k} \in \Theta_{-k}} \|\mathbf{h}_{kn}(\boldsymbol{\theta}_{-k}) - \mathbf{h}_k^*(\boldsymbol{\theta}_{-k})\| = O_p(\epsilon_n)$, for all $k = 1, \dots, K$.

(A3) The sequence of maps $\{\Psi_n\}_{n \geq 1}$ is continuous in $\boldsymbol{\lambda}_{1:K} \in \mathbf{\Lambda}_{1:K}$. Moreover, the sequence of solution set $\mathcal{S}_n$ is non-empty almost surely.

(A4) There exists a unique $\boldsymbol{\lambda}_{1:K}^* \in \mathbf{\Lambda}_{1:K}$ that satisfies

$$\int \mathbf{T}_k(\boldsymbol{\theta}_k) q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^*) d\boldsymbol{\theta}_k = \int^{\otimes^{K-1}} \mathbf{h}_k^*(\boldsymbol{\theta}_{-k}) \prod_{j \neq k} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j^*) d\boldsymbol{\theta}_j,$$

for all $k = 1, \dots, K$. The integral $\int \mathbf{T}_k(\boldsymbol{\theta}_k) q_k(\boldsymbol{\theta}_k; \cdot) d\boldsymbol{\theta}_k$ is continuous in $\boldsymbol{\lambda}_k \in \mathbf{\Lambda}_k$ and $\int^{\otimes^{K-1}} \mathbf{h}_1^*(\boldsymbol{\theta}_{-1}) \prod_{j \neq 1} q(\boldsymbol{\theta}_j; \cdot) d\boldsymbol{\theta}_j$ is continuous in $\boldsymbol{\lambda}_{-1} \in \mathbf{\Lambda}_{-1}$.

(A5) Denote

$$\Psi(\boldsymbol{\lambda}) = \mathbf{m}(\boldsymbol{\lambda}) - \mathbf{G}(\boldsymbol{\lambda})$$

where

$$\mathbf{G}(\boldsymbol{\lambda}) = \begin{pmatrix} \int^{\otimes^{K-1}} \mathbf{h}_1^*(\boldsymbol{\theta}_{-1}) \prod_{j \neq 1} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j \\ \vdots \\ \int^{\otimes^{K-1}} \mathbf{h}_K^*(\boldsymbol{\theta}_{-K}) \prod_{j \neq K} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j \end{pmatrix}.$$

For every sufficiently large $\omega > 0$, the following lower bound holds: $\inf_{\boldsymbol{\lambda}_{1:K} \in \mathcal{G}_n(\omega)} \|\Psi(\boldsymbol{\lambda})\| \geq M_\omega \epsilon_n$, where $\mathcal{G}_n(\omega) = \{\boldsymbol{\lambda}_{1:K} \in \mathbf{\Lambda}_{1:K} : \|\boldsymbol{\lambda}_{1:K} - \boldsymbol{\lambda}_{1:K}^*\| \geq \omega \epsilon_n\}$, $\boldsymbol{\lambda}_{1:K}^*$ is a point in $\mathbf{\Lambda}_{1:K}$ such that $\Psi(\boldsymbol{\lambda}_{1:K}^*) = \mathbf{0}$, and $\lim_{\omega \rightarrow \infty} M_\omega = \infty$.

(A6) $\boldsymbol{\mu}_k(\boldsymbol{\lambda}_k^*) = \boldsymbol{\theta}_k^*$ and $\mathbf{U}_k(\boldsymbol{\lambda}_k^*) = \boldsymbol{\theta}_k^* \boldsymbol{\theta}_k^{*\top}$ for all $k = 1, \dots, K$.

(A7) The marginal posterior distribution of the localized parameter $\mathbf{z}_k = \sqrt{n}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^*)$ converges in total variation distance to a centered Gaussian distribution:

$$\text{DTV}\{p(\mathbf{z}_k \mid \mathcal{D}_n) \parallel \text{MVN}(\mathbf{0}, \mathbf{V}_k^*)\} \xrightarrow{\mathbb{P}^*} 0$$

where $\mathbf{V}_k^*$ is a strictly positive definite asymptotic covariance matrix. Furthermore, we assume the scaled posterior covariance converges: $n \text{Var}(\boldsymbol{\theta}_k \mid \mathcal{D}_n) \xrightarrow{\mathbb{P}^*} \mathbf{V}_k^*$.

(A8) The full conditional mean $\mathbf{m}_{kn}$ is continuously differentiable with a bounded Jacobian

matrix in $\mathcal{B}_{-k} = \{\boldsymbol{\theta}_{-k} : \|\boldsymbol{\theta}_{-k} - \boldsymbol{\theta}_{-k}^*\| \leq R\}$. The full conditional variance is bounded $\|\boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k})\| \leq M_{\text{prior}}$ for all $\boldsymbol{\theta}_{-k} \in \boldsymbol{\Theta}_{-k}$. Moreover, for all $\boldsymbol{\theta}_{-k} \in \mathcal{B}_{-k}$, the following expansion holds: $\boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k}) = n^{-1}\tilde{\boldsymbol{\omega}}_{kn}(\boldsymbol{\theta}_{-k}) + \mathbf{r}_n(\boldsymbol{\theta}_{-k})$. The function $\tilde{\boldsymbol{\omega}}_{kn}$ is strictly positive definite and twice-continuously differentiable with a bounded Hessian matrix, and the remainder satisfies $\sup_{\boldsymbol{\theta}_{-k} \in \mathcal{B}_{-k}} \|n \cdot \mathbf{r}_n(\boldsymbol{\theta}_{-k})\| \rightarrow 0$ almost surely.

- (A9) For each n , the pair of functions $\boldsymbol{\lambda}_k \mapsto (\boldsymbol{\mu}_k(\boldsymbol{\lambda}_k), \boldsymbol{\Gamma}_k^{(n)}(\boldsymbol{\lambda}_k))$ defines a unique mapping between the variational parameter space $\boldsymbol{\Lambda}_k$ and the location-variance space. Furthermore, this mapping is a twice-continuously differentiable diffeomorphism in an open neighborhood of $\boldsymbol{\lambda}_k^*$ with a nonsingular Jacobian. Moreover, the limit $\boldsymbol{\Gamma}_k^* := \lim_{n \rightarrow \infty} \boldsymbol{\Gamma}_k^{(n)}(\hat{\boldsymbol{\lambda}}_k)$ is strictly positive definite and has bounded eigenvalues.

Condition (A1) is necessary to guarantee the existence of a solution under most versions of fixed point theorems. This assumption required to establish uniform convergence of the finite n fixed point conditions to their respective limits. Condition (A2) assumes the exact posterior is consistent at rate ϵ_n and the expectations of the full conditionals are uniformly convergent. Condition (A3) is guarantees the existence of a sequence solutions to the finite n fixed point equations. Condition (A4) assumes the uniqueness of solution to the limiting fixed point problem. Condition (A5) is an identifiability condition on the unique solution to the limiting fixed point equations. Condition (A6) is required to ensure that the solution to the limiting MFMP system of integral equations corresponds to a dirac mass on the true value. Condition (A7) ensures that the individual K exact marginal posteriors exhibit a Bernstein-von Mises property and is required to establish asymptotic normality of the MFMP posterior. Condition (A8) is a smoothness assumption on the full conditional mean and variances. Condition (A9) is a smoothness assumption on the mapping between the variational parameters and the location-scale space. (A8) and (A9) are essential for proving asymptotic equivalence between the MFMP posterior variance and the exact posterior variance. Condition (A10) assumes that the choice of approximating distributions admit an asymptotically normal form as n diverges.

B.1 Statement and proof of Theorem 1

Assume (A1) to (A5) holds. Then, there exists a sequence of $\{n_l\}_{l=1}^\infty$ such that $\mathcal{S}_{n_l} \neq \emptyset$.

Moreover, there exists a corresponding sequence of solutions $\{\hat{\boldsymbol{\lambda}}_{1:K}^{(n_\ell)}\}_{\ell=1}$ such that

$$\|\hat{\boldsymbol{\lambda}}_{1:K}^{(n_\ell)} - \boldsymbol{\lambda}_{1:K}^*\| = O_p(\epsilon_n).$$

Proof 1 The existence of an increasing sequence $\{n_l\}_{l=1}^\infty$ such that $\mathcal{S}_{n_l} \neq \emptyset$ follows from (A3).

Consider the event $\{\hat{\boldsymbol{\lambda}}_{1:K}^{(n_l)} \in \mathcal{G}_{n_l}(\omega)\}$. Then, since $\boldsymbol{\Psi}_{n_l}(\hat{\boldsymbol{\lambda}}_{1:K}^{(n_l)}) = \mathbf{0}$, the event $\{\hat{\boldsymbol{\lambda}}_{1:K}^{(n_l)} \in \mathcal{G}_{n_l}(\omega)\}$ implies the event

$$\{\|\boldsymbol{\Psi}(\hat{\boldsymbol{\lambda}}_{1:K}^{(n_l)}) - \boldsymbol{\Psi}_{n_l}(\hat{\boldsymbol{\lambda}}_{1:K}^{(n_l)})\| \geq M_\omega \epsilon_{n_l}\}$$

Then,

$$\begin{aligned} \mathbb{P}^* \left\{ \hat{\boldsymbol{\lambda}}_{1:K}^{(n_l)} \in \mathcal{G}_{n_l}(\omega) \right\} &\leq \mathbb{P}^* \left\{ \|\boldsymbol{\Psi}(\hat{\boldsymbol{\lambda}}_{1:K}^{(n_l)}) - \boldsymbol{\Psi}_{n_l}(\hat{\boldsymbol{\lambda}}_{1:K}^{(n_l)})\| \geq M_\omega \epsilon_{n_l} \right\} \\ &\leq \mathbb{P}^* \left\{ \sup_{\boldsymbol{\lambda}_{1:K} \in \boldsymbol{\Lambda}_{1:K}} \|\boldsymbol{\Psi}(\boldsymbol{\lambda}_{1:K}) - \boldsymbol{\Psi}_{n_l}(\boldsymbol{\lambda}_{1:K})\| \geq M_\omega \epsilon_{n_l} \right\}. \end{aligned}$$

From Lemma 1, for any $\delta > 0$, we can find a (ω, M_ω) pairing such that

$$\mathbb{P}^* \left\{ \hat{\boldsymbol{\lambda}}_{1:K}^{(n_l)} \in \mathcal{G}_{n_l}(\omega) \right\} \leq \mathbb{P}^* \left\{ \sup_{\boldsymbol{\lambda}_{1:K} \in \boldsymbol{\Lambda}_{1:K}} \|\boldsymbol{\Psi}(\boldsymbol{\lambda}_{1:K}) - \boldsymbol{\Psi}_{n_l}(\boldsymbol{\lambda}_{1:K})\| \geq M_\omega \epsilon_{n_l} \right\} < \delta.$$

Hence, $\|\hat{\boldsymbol{\lambda}}_{1:K} - \boldsymbol{\lambda}_{1:K}^*\| = O_p(\epsilon_n)$ as required.

B.2 Statement and proof of Theorem 2

Assume (A1) to (A6) holds. Then, for any $\epsilon > 0$,

$$\mathbb{P}_{q_k(\cdot; \hat{\boldsymbol{\lambda}}_k)} \{ \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^*\| > \epsilon \} \xrightarrow{\mathbb{P}^*} 0.$$

Proof 2 By the continuity of $\int \boldsymbol{\theta}_k q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k) d\boldsymbol{\theta}_k$ and $\int \boldsymbol{\theta}_k \boldsymbol{\theta}_k^\top q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k) d\boldsymbol{\theta}_k$ in $\boldsymbol{\lambda}_k \in \boldsymbol{\Lambda}_k$ as

assumed in (A4) and Theorem 1, we have (by CMT):

$$\boldsymbol{\mu}_k(\widehat{\boldsymbol{\lambda}}_k) \xrightarrow{\mathbb{P}^*} \int \boldsymbol{\theta}_k q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^*) d\boldsymbol{\theta}_k = \boldsymbol{\theta}_k^*$$

and

$$\mathbf{U}_k(\widehat{\boldsymbol{\lambda}}_k) \xrightarrow{\mathbb{P}^*} \int \boldsymbol{\theta}_k \boldsymbol{\theta}_k^\top q_k(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k^*) d\boldsymbol{\theta}_k = \boldsymbol{\theta}_k^* \boldsymbol{\theta}_k^{*\top}.$$

Hence

$$\frac{1}{n} \boldsymbol{\Gamma}_k^{(n)}(\widehat{\boldsymbol{\lambda}}_k) \xrightarrow{\mathbb{P}^*} \mathbf{0}.$$

Then,

$$\mathbb{P}_{q_k(\cdot; \widehat{\boldsymbol{\lambda}}_k)} \{ \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^*\| > \epsilon \} \leq \mathbb{P}_{q_k(\cdot; \widehat{\boldsymbol{\lambda}}_k)} \{ \|\boldsymbol{\theta}_k - \boldsymbol{\mu}_k(\widehat{\boldsymbol{\lambda}}_k)\| > \epsilon/2 \} + \mathbb{I} \{ \|\boldsymbol{\mu}_k(\widehat{\boldsymbol{\lambda}}_k) - \boldsymbol{\theta}_k^*\| > \epsilon/2 \}.$$

Now, by Chebychev's inequality, we have

$$\mathbb{P}_{q_k(\cdot; \widehat{\boldsymbol{\lambda}}_k)} \{ \|\boldsymbol{\theta}_k - \boldsymbol{\mu}_k(\widehat{\boldsymbol{\lambda}}_k)\| > \epsilon/2 \} \leq \frac{4 \text{tr}(\mathbb{V}_{\boldsymbol{\theta}_k \sim q_k(\cdot; \widehat{\boldsymbol{\lambda}}_k)}[\boldsymbol{\theta}_k])}{\epsilon^2} \xrightarrow{\mathbb{P}^*} 0.$$

Moreover, by (A6), we have

$$\|\boldsymbol{\mu}_k(\widehat{\boldsymbol{\lambda}}_k) - \boldsymbol{\theta}_k^*\| = \|\boldsymbol{\mu}_k(\widehat{\boldsymbol{\lambda}}_k) - \boldsymbol{\mu}_k(\boldsymbol{\lambda}_k^*)\| \xrightarrow{\mathbb{P}^*} 0,$$

where the last stochastic convergence holds by CMT. Hence, we have proved our required result.

B.3 Statement and proof of Theorem 3

Assume (A1) to (A9) holds. Then, for each $k = 1, 2$, we have

$$\|\boldsymbol{\Sigma}_{kn} - \mathbf{V}_{kn}\| = O_p(n^{-3/2}),$$

where $\mathbf{V}_{kn} = \mathbb{V}_{p(\boldsymbol{\theta}_k | \mathcal{D}_n)}(\boldsymbol{\theta}_k | \mathcal{D}_n)$.

Proof 3 Consider the sequence of MFMP solutions $\{\widehat{\boldsymbol{\lambda}}_k^{(n)}\}_{n \geq 1}$ and their corresponding location-

scale map

$$\{(\boldsymbol{\mu}_k(\widehat{\boldsymbol{\lambda}}_k^{(n)}), \boldsymbol{\Gamma}_k^{(n)}(\widehat{\boldsymbol{\lambda}}_k^{(n)}))\}_{n \geq 1}.$$

For notational brevity, we will write shorthand $\boldsymbol{\mu}_{kn} := \boldsymbol{\mu}_k(\widehat{\boldsymbol{\lambda}}_k^{(n)})$, $\boldsymbol{\Gamma}_{kn} := \boldsymbol{\Gamma}_k^{(n)}(\widehat{\boldsymbol{\lambda}}_k^{(n)})$, and $\boldsymbol{\Sigma}_{kn} := \frac{1}{n} \boldsymbol{\Gamma}_k^{(n)}(\widehat{\boldsymbol{\lambda}}_k^{(n)})$. Now, by definition of the MFMP solution

$$\begin{aligned} \boldsymbol{\Sigma}_{kn} &= \text{Var}_{\tilde{r}_k(\cdot; \widehat{\boldsymbol{\lambda}}_{-k}^{(n)})}(\boldsymbol{\theta}_k) \\ &= \mathbb{E}_{q_{-k}}[\boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k})] + \text{Var}_{q_{-k}}[\mathbf{m}_{kn}(\boldsymbol{\theta}_{-k})]. \end{aligned}$$

where we recall that $\tilde{r}_k(\boldsymbol{\theta}_{-k}; \widehat{\boldsymbol{\lambda}}_{-k}^{(n)}) = \int p(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{-k}, \mathcal{D}_n) q(\boldsymbol{\theta}_{-k}; \widehat{\boldsymbol{\lambda}}_{-k}^{(n)}) d\boldsymbol{\theta}_{-k}$. We decompose the integral $\mathbb{E}_{q_{-k}}[\boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k})]$ using neighbourhood $\mathcal{B}_{-k}$:

$$\mathbb{E}_{q_{-k}}[\boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k})] = \int_{\mathcal{B}_{-k}} \boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k}) q(\boldsymbol{\theta}_{-k}; \widehat{\boldsymbol{\lambda}}_{-k}^{(n)}) d\boldsymbol{\theta}_{-k} + \int_{\mathcal{B}_{-k}^c} \boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k}) q(\boldsymbol{\theta}_{-k}; \widehat{\boldsymbol{\lambda}}_{-k}^{(n)}) d\boldsymbol{\theta}_{-k}. \quad (25)$$

Due to the upper bound of $\boldsymbol{\omega}_{kn}$ from (A8), we can write

$$\begin{aligned} \int_{\mathcal{B}_{-k}^c} \boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} &\leq M_{\text{prior}} \int_{\mathcal{B}_{-k}^c} q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \\ &\leq \frac{M_{\text{prior}} \mathbb{E} \left[\|\boldsymbol{\theta}_k - \boldsymbol{\mu}_k(\widehat{\boldsymbol{\lambda}}_k^{(n)})\|^4 \right]}{(R/2)^4} \\ &= O_p(n^{-2}) \\ &= o_p(n^{-1}), \end{aligned}$$

where we write $q_{-k}(\boldsymbol{\theta}_{-k})$ in place of $q_{-k}(\boldsymbol{\theta}_{-k}; \widehat{\boldsymbol{\lambda}}_{-k}^{(n)})$ for notational brevity. Hence

$$\mathbb{E}_{q_{-k}}[\boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k})] = \int_{\mathcal{B}_{-k}} \boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} + o_p(n^{-1}).$$

To complete our analysis of $\mathbb{E}_{q_{-k}}[\boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k})]$, it remains for us to analyse the first term in the RHS of equation (25). In fact, by (A8),

$$\begin{aligned} &\int_{\mathcal{B}_{-k}} \boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \\ &\leq n^{-1} \int_{\mathcal{B}_{-k}} \tilde{\boldsymbol{\omega}}_{kn}(\boldsymbol{\theta}_{-k}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} + \int_{\mathcal{B}_{-k}} \mathbf{r}_k(\boldsymbol{\theta}_{-k}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \end{aligned}$$

Now,

$$\left\| \int_{\mathcal{B}_{-k}} \mathbf{r}_k(\boldsymbol{\theta}_{-k}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \right\| \leq \int_{\mathcal{B}_{-k}} \|\mathbf{r}_k(\boldsymbol{\theta}_{-k})\| q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} = o_p(n^{-1}),$$

where the last stochastic convergence equality follows from (A8). To understand the asymptotic behaviour of the term $n^{-1} \int_{\mathcal{B}_{-k}} \tilde{\omega}_k(\boldsymbol{\theta}_{-k}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k}$, we consider an elementwise second-order Taylor's expansion of $\tilde{\omega}_{kn}$. In particular,

$$\begin{aligned} \tilde{\omega}_{knj}(\boldsymbol{\theta}_{-k}) &= \tilde{\omega}_{knj}(\boldsymbol{\mu}_{-kn}) + \nabla_{\boldsymbol{\theta}_{-k}} \tilde{\omega}_{knj}(\boldsymbol{\mu}_{-kn})(\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}) \\ &\quad + \frac{1}{2}(\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn})^\top \frac{\partial^2 \tilde{\omega}_{knj}}{\partial \boldsymbol{\theta}_{-k} \partial \boldsymbol{\theta}_{-k}^\top}(\bar{\boldsymbol{\theta}}_{-k})(\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}), \end{aligned}$$

where we use the shorthand notation $\boldsymbol{\mu}_{-kn} := \int \boldsymbol{\theta}_{-k} q(\boldsymbol{\theta}_{-k}; \hat{\boldsymbol{\lambda}}_{-k}^{(n)}) d\boldsymbol{\theta}_{-k}$. Hence, for each entry $\tilde{\omega}_{knj}(\boldsymbol{\theta}_{-k})$, we have

$$\begin{aligned} &\int_{\mathcal{B}_{-k}} \tilde{\omega}_{knj}(\boldsymbol{\theta}_{-k}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \\ &= \tilde{\omega}_{knj}(\boldsymbol{\mu}_{-kn}) + \nabla_{\boldsymbol{\theta}_{-k}} \tilde{\omega}_{knj}(\boldsymbol{\mu}_{-kn}) \int_{\mathcal{B}_{-k}} (\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \\ &\quad + \frac{1}{2} \int_{\mathcal{B}_{-k}} (\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn})^\top \frac{\partial^2 \tilde{\omega}_{knj}}{\partial \boldsymbol{\theta}_{-k} \partial \boldsymbol{\theta}_{-k}^\top}(\bar{\boldsymbol{\theta}}_{-k})(\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \end{aligned}$$

Now,

$$\mathbf{0} = \int_{\mathcal{B}_{-k}} (\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} + \int_{\mathcal{B}_{-k}^c} (\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k}$$

and hence

$$\left\| \int_{\mathcal{B}_{-k}} (\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \right\| = \left\| \int_{\mathcal{B}_{-k}^c} (\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \right\|.$$

Following Cauchy-Schwarz inequality, we show that

$$\begin{aligned} &\left\| \int_{\mathcal{B}_{-k}^c} (\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \right\| \\ &\leq \int_{\boldsymbol{\Theta}_{-k}} \|\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}\| \mathbb{I}\{\boldsymbol{\theta}_{-k} \in \mathcal{B}_{-k}^c\} q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \\ &\leq \sqrt{\mathbb{E}_{q_{-k}} [\|\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}\|^2]} \times \underbrace{\sqrt{\int_{\mathcal{B}_{-k}^c} q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k}}}_{\leq \frac{\mathbb{E}[\|\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}\|^4]}{(R/2)^4}} = o_p(n^{-2}). \end{aligned}$$

Following (A8), each $\tilde{\omega}_{knj}$ is twice continuously differentiable. Hence there exists a constant M such that $\left\| \frac{\partial^2 \tilde{\omega}_{knj}}{\partial \boldsymbol{\theta}_{-k} \partial \boldsymbol{\theta}_{-k}^\top}(\boldsymbol{\theta}_{-k}) \right\| \leq M$ for all $\boldsymbol{\theta}_{-k} \in \boldsymbol{\Theta}_{-k}$. Thus,

$$\begin{aligned} & \int_{\mathcal{B}_{-k}} (\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn})^\top \frac{\partial^2 \tilde{\omega}_{knj}}{\partial \boldsymbol{\theta}_{-k} \partial \boldsymbol{\theta}_{-k}^\top}(\bar{\boldsymbol{\theta}}_{-k}) (\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}) d\boldsymbol{\theta}_{-k} \\ & \leq M \int_{\mathcal{B}_{-k}} \|\boldsymbol{\theta}_{-k} - \boldsymbol{\mu}_{-kn}\|^2 q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} \leq M \text{tr}(\boldsymbol{\Sigma}_{-kn}) = O_p(n^{-1}), \end{aligned}$$

where the last stochastic convergence follows from an implication on (A9), i.e., $\|\boldsymbol{\Sigma}_{-kn}\| = O_p(n^{-1})$. Hence, we have

$$\int_{\mathcal{B}_{-k}} \boldsymbol{\omega}_{kn}(\boldsymbol{\theta}_{-k}) q_{-k}(\boldsymbol{\theta}_{-k}) d\boldsymbol{\theta}_{-k} = n^{-1} \tilde{\boldsymbol{\omega}}_{kn}(\boldsymbol{\mu}_{-kn}) + o_p(n^{-1}).$$

Furthermore, by delta method, it is straightforward to show that

$$\mathbb{V}_{q_{-k}}[\mathbf{m}_{kn}(\boldsymbol{\theta}_{-k})] = \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\mu}_{-kn})^\top \boldsymbol{\Sigma}_{-kn} \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\mu}_{-kn}) + o_p(n^{-1}).$$

Moreover, since $\boldsymbol{\Sigma}_{-kn} = O_p(n^{-1})$, hence $\text{Var}_{q_{-k}}[\mathbf{m}_{kn}(\boldsymbol{\theta}_{-k})] = O_p(n^{-1})$. Thus, we have shown that

$$n \boldsymbol{\Sigma}_{kn} = \tilde{\boldsymbol{\omega}}_{kn}(\boldsymbol{\mu}_{-kn}) + \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\mu}_{-kn})^\top \boldsymbol{\Sigma}_{-kn} \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\mu}_{-kn}) + o_p(1). \quad (26)$$

Next, we analyse the asymptotic behaviour of the exact posterior variance $\mathbf{V}_{kn}$. By similar application of Taylor's expansion and delta method in the preceding segment of this proof, we can also show that

$$n \mathbf{V}_{kn} = \tilde{\boldsymbol{\omega}}_{kn}(\boldsymbol{\theta}_{-k}^*) + \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\theta}_{-k}^*)^\top \mathbf{V}_{-kn} \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\theta}_{-k}^*) + o_p(1). \quad (27)$$

Hence, by combining (26) and (27), we have

$$\begin{aligned} & n \|\boldsymbol{\Sigma}_{kn} - \mathbf{V}_{kn}\| \\ & \leq \|\tilde{\boldsymbol{\omega}}_{kn}(\boldsymbol{\mu}_{-kn}) - \tilde{\boldsymbol{\omega}}_{kn}(\boldsymbol{\theta}_{-k}^*)\| \\ & + |\nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\mu}_{-kn})^\top \boldsymbol{\Sigma}_{-kn} \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\mu}_{-kn}) - \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\theta}_{-k}^*)^\top \boldsymbol{\Sigma}_{-kn} \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\theta}_{-k}^*)|. \end{aligned} \quad (28)$$

To bound the first term on the RHS of (28), we note from (A8) that $\tilde{\omega}_{kn}$ is twice continuously-differentiable and Lemma 2 that $\|\mu_{-kn} - \theta_{-k}^*\| = O_p(n^{-1/2})$. Hence

$$\begin{aligned} & \|\tilde{\omega}_{kn}(\mu_{-kn}) - \tilde{\omega}_{kn}(\theta_{-k}^*)\| \\ & \leq \underbrace{\sup_{t \in [0,1]} \left\| \nabla_{\theta_{-k}} \tilde{\omega}_{kn}(t\mu_{-kn} + (1-t)\theta_{-k}^*) \right\|}_{=O_p(1)} \|\mu_{-kn} - \theta_{-k}^*\| = O_p(n^{-1/2}). \end{aligned}$$

Moreover, we can bound the second term in (28) as

$$\begin{aligned} & |\nabla_{\theta_{-k}} \mathbf{m}_{kn}(\mu_{-kn})^\top \Sigma_{-kn} \nabla_{\theta_{-k}} \mathbf{m}_{kn}(\mu_{-kn}) - \nabla_{\theta_{-k}} \mathbf{m}_{kn}(\theta_{-k}^*)^\top \Sigma_{-kn} \nabla_{\theta_{-k}} \mathbf{m}_{kn}(\theta_{-k}^*)| \\ & \leq \|\nabla_{\theta_{-k}} \mathbf{m}_{kn}(\mu_{-kn}) - \nabla_{\theta_{-k}} \mathbf{m}_{kn}(\theta_{-k}^*)\| \|\Sigma_{-kn}\| \underbrace{\|\nabla_{\theta_{-k}} \mathbf{m}_{kn}(\mu_{-kn}) + \nabla_{\theta_{-k}} \mathbf{m}_{kn}(\theta_{-k}^*)\|}_{=O_p(1)} \\ & + \|\nabla_{\theta_{-k}} \mathbf{m}_{kn}(\theta_{-k}^*)\|^2 \|\Sigma_{-kn} - \mathbf{V}_{-kn}\| \end{aligned}$$

Following the twice-continuous differentiability of $\mathbf{m}_{kn}$ in (A8) and Lemma 2, we have

$$\begin{aligned} & \|\nabla_{\theta_{-k}} \mathbf{m}_{kn}(\mu_{-kn}) - \nabla_{\theta_{-k}} \mathbf{m}_{kn}(\theta_{-k}^*)\| \\ & \leq \underbrace{\sup_{t \in [0,1]} \left\| \nabla_{\theta_{-k}} \mathbf{m}_{kn}(t\mu_{-kn} + (1-t)\theta_{-k}^*) \right\|}_{=O_p(1)} \|\mu_{-kn} - \theta_{-k}^*\| = O_p(n^{-1/2}). \end{aligned}$$

Moreover, recall from Lemma 2 that $\|\Sigma_{-kn}\| = O_p(n^{-1})$. Now, by following similar steps as above, we obtain

$$\begin{aligned} & n \|\Sigma_{-kn} - \mathbf{V}_{-kn}\| \\ & \leq O_p(n^{-1/2}) + \|\nabla_{\theta_k} \mathbf{m}_{-kn}(\mu_{kn}) - \nabla_{\theta_k} \mathbf{m}_{-kn}(\theta_k^*)\| \|\Sigma_{kn}\| \|\nabla_{\theta_k} \mathbf{m}_{-kn}(\mu_{kn}) + \nabla_{\theta_k} \mathbf{m}_{-kn}(\theta_k^*)\| \\ & + \|\nabla_{\theta_k} \mathbf{m}_{-kn}(\theta_k^*)\|^2 \|\Sigma_{kn} - \mathbf{V}_{kn}\| \end{aligned} \tag{29}$$

By (A7) and Lemma 2, we have $\|\Sigma_{kn}\| = O_p(n^{-1})$ and $\|\mathbf{V}_{kn}\| = O_p(n^{-1})$. Hence, a crude stochastic convergence order for $\|\Sigma_{kn} - \mathbf{V}_{kn}\|$ is $O_p(n^{-1})$. Hence, $\|\Sigma_{-kn} - \mathbf{V}_{-kn}\| = O_p(n^{-3/2})$.

Consequently,

$$|\nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\mu}_{-kn})^\top \boldsymbol{\Sigma}_{-kn} \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\mu}_{-kn}) - \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\theta}_{-k}^*)^\top \boldsymbol{\Sigma}_{-kn} \nabla_{\boldsymbol{\theta}_{-k}} \mathbf{m}_{kn}(\boldsymbol{\theta}_{-k}^*)| = O_p(n^{-3/2}).$$

Hence, substituting the previous equation back into (28), we have

$$\|\boldsymbol{\Sigma}_{kn} - \mathbf{V}_{kn}\| = O_p(n^{-3/2})$$

as required.

B.4 Supporting Lemmas

Lemma 1 Assume (A1) and (A2) holds. Then,

$$\sup_{\boldsymbol{\lambda}_{1:K} \in \boldsymbol{\Lambda}_{1:K}} \|\boldsymbol{\Psi}_n(\boldsymbol{\lambda}) - \boldsymbol{\Psi}(\boldsymbol{\lambda})\| = O_p(\epsilon_n).$$

Proof 4 Denote

$$\boldsymbol{\Psi}(\boldsymbol{\lambda}) = \mathbf{m}(\boldsymbol{\lambda}) - \mathbf{G}(\boldsymbol{\lambda})$$

where

$$\mathbf{G}(\boldsymbol{\lambda}) = \begin{pmatrix} \int^{\otimes K-1} \mathbf{h}_1^*(\boldsymbol{\theta}_{-1}) \prod_{j \neq 1} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j \\ \vdots \\ \int^{\otimes K-1} \mathbf{h}_K^*(\boldsymbol{\theta}_{-K}) \prod_{j \neq K} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j \end{pmatrix}.$$

For any $\boldsymbol{\lambda}_{1:K} \in \boldsymbol{\Lambda}_{1:K}$,

$$\begin{aligned} \|\boldsymbol{\Psi}_n(\boldsymbol{\lambda}_{1:K}) - \boldsymbol{\Psi}(\boldsymbol{\lambda}_{1:K})\| &= \|\mathbf{G}_n(\boldsymbol{\lambda}) - \mathbf{G}(\boldsymbol{\lambda})\| \\ &\leq \sum_{k=1}^K \int \|\mathbf{h}_{kn}(\boldsymbol{\theta}_{-k}) - \mathbf{h}_k^*(\boldsymbol{\theta}_{-k})\| \prod_{j \neq k} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j \\ &\leq \sum_{k=1}^K \sup_{\boldsymbol{\theta}_{-k} \in \boldsymbol{\Theta}_{-k}} \|\mathbf{h}_{kn}(\boldsymbol{\theta}_{-k}) - \mathbf{h}_k^*(\boldsymbol{\theta}_{-k})\| \underbrace{\int \prod_{j \neq k} q(\boldsymbol{\theta}_j; \boldsymbol{\lambda}_j) d\boldsymbol{\theta}_j}_{=1} \\ &= \sum_{k=1}^K \sup_{\boldsymbol{\theta}_{-k} \in \boldsymbol{\Theta}_{-k}} \|\mathbf{h}_{kn}(\boldsymbol{\theta}_{-k}) - \mathbf{h}_k^*(\boldsymbol{\theta}_{-k})\|, \end{aligned}$$

where the first equality follows from the definition of Ψ_n and Ψ , the second inequality follows from the definition of $\mathbf{G}_n$ and $\mathbf{G}$ and then an application of triangle inequality, the third inequality follows from taking sup over $\boldsymbol{\theta}_{-k}$. By noting from (A1) that $\boldsymbol{\Lambda}_{1:K}$ is compact and the above upper bound doesn't depend on $\boldsymbol{\lambda}_{1:K}$, then

$$\sup_{\boldsymbol{\lambda}_{1:K} \in \boldsymbol{\Lambda}_{1:K}} \|\Psi_n(\boldsymbol{\lambda}_{1:K}) - \Psi(\boldsymbol{\lambda}_{1:K})\| \leq \sum_{k=1}^K \sup_{\boldsymbol{\theta}_{-k} \in \boldsymbol{\Theta}_{-k}} \|\mathbf{h}_{kn}(\boldsymbol{\theta}_{-k}) - \mathbf{h}_k^*(\boldsymbol{\theta}_{-k})\| = O_p(\epsilon_n),$$

where the stochastic convergence order follows from (A2).

Lemma 2 Assume (A1) to (A6) and (A9) holds. Then, for each $k = 1, \dots, K$, we have

$$\|\boldsymbol{\mu}_k(\hat{\boldsymbol{\lambda}}_k) - \boldsymbol{\theta}_k^*\| = O_p(n^{-1/2})$$

and

$$\|\frac{1}{n} \boldsymbol{\Gamma}_k^{(n)}(\hat{\boldsymbol{\lambda}}_k)\| = O_p(n^{-1}).$$

Proof 5 Following Theorem 1, our sequence of solutions converges as $\|\hat{\boldsymbol{\lambda}}_k - \boldsymbol{\lambda}_k^*\| = O_p(n^{-1/2})$. Following a straightforward Taylor's expansion, and applying the convergence rate of $\|\hat{\boldsymbol{\lambda}}_k - \boldsymbol{\lambda}_k^*\|$, we have

$$\|\boldsymbol{\mu}_k(\hat{\boldsymbol{\lambda}}_k) - \boldsymbol{\mu}_k(\boldsymbol{\lambda}_k^*)\| \leq \|\nabla_{\boldsymbol{\lambda}_k} \boldsymbol{\mu}_k(\boldsymbol{\lambda}_k^*)\| \|\hat{\boldsymbol{\lambda}}_k - \boldsymbol{\lambda}_k^*\| + o_p(n^{-1/2})$$

and

$$\|\text{vec}(\boldsymbol{\Gamma}_k^{(n)}(\hat{\boldsymbol{\lambda}}_k)) - \text{vec}(\boldsymbol{\Gamma}_k^{(n)}(\boldsymbol{\lambda}_k^*))\| \leq \|\nabla_{\boldsymbol{\lambda}_k} \text{vec}(\boldsymbol{\Gamma}_k^{(n)}(\boldsymbol{\lambda}_k^*))\| \|\hat{\boldsymbol{\lambda}}_k - \boldsymbol{\lambda}_k^*\| + o_p(n^{-1/2}).$$

From (A9), the map $\boldsymbol{\lambda}_k \mapsto (\boldsymbol{\mu}_k(\boldsymbol{\lambda}_k), \boldsymbol{\Gamma}_k^{(n)}(\boldsymbol{\lambda}_k))$ is twice-continuous differentiability. Hence, the two matrix norms of the first-order derivative are bounded: $\|\nabla_{\boldsymbol{\lambda}_k} \boldsymbol{\mu}_k(\boldsymbol{\lambda}_k^*)\| = O(1)$ and $\|\nabla_{\boldsymbol{\lambda}_k} \text{vec}(\boldsymbol{\Gamma}_k^{(n)}(\boldsymbol{\lambda}_k^*))\| = O(1)$. Hence, by (A6), we have

$$\|\boldsymbol{\mu}_k(\hat{\boldsymbol{\lambda}}_k) - \boldsymbol{\mu}_k(\boldsymbol{\lambda}_k^*)\| = \|\boldsymbol{\mu}_k(\hat{\boldsymbol{\lambda}}_k) - \boldsymbol{\theta}_k^*\| = O_p(n^{-1/2})$$

and

$$\|\text{vec}(\boldsymbol{\Gamma}_k^{(n)}(\hat{\boldsymbol{\lambda}}_k)) - \text{vec}(\boldsymbol{\Gamma}_k^{(n)}(\boldsymbol{\lambda}_k^*))\| = O_p(n^{-1/2}).$$

Thus have

$$\|vec(\frac{1}{n}\mathbf{\Gamma}_k^{(n)}(\hat{\boldsymbol{\lambda}}_k))\| \leq \frac{1}{n}\|vec(\mathbf{\Gamma}_k^{(n)}(\hat{\boldsymbol{\lambda}}_k)) - vec(\mathbf{\Gamma}_k^{(n)}(\boldsymbol{\lambda}_k^*))\| + \frac{1}{n}\|vec(\mathbf{\Gamma}_k^{(n)}(\boldsymbol{\lambda}_k^*))\| = O_p(n^{-1}).$$

C Implementation details for GMP

This section gives implementation details for Gaussian moment propagation. Section C.1 shows how the local quantities required for $r_k(\theta_k)$ can be computed efficiently from precomputed quantities $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\mu} \in \mathbb{R}^n$, $\mathbf{M} = \mathbf{X}\boldsymbol{\Sigma} \in \mathbb{R}^{n \times p}$, and $\mathbf{s}^2 = \text{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T) \in \mathbb{R}^n$. Section C.2 derives the Gaussian joint update after replacing a marginal distribution. Section C.3 proves positive definiteness of the resulting covariance matrix. Section C.4 shows how to update the precomputed quantities $\boldsymbol{\eta}$, $\mathbf{M}$, and $\mathbf{s}^2$ after each coordinate update. This reduces the cost of the updates and dwnupdates from $O(np^2)$ to $O(np + p^2)$ for each coordinate k .

For simplicity in Section C.1 and Section C.4, because this is the case used in Sections 4.5 and 4.6, we describe the implementation for scalar coordinate updates. Block updates are analogous but require block-matrix versions of the same formulae.

C.1 Efficient dwnupdates

The calculation of (17) involves the quantities

- $\mathbf{a} = \boldsymbol{\Sigma}_{-k,k} / \Sigma_{k,k}$,
- $\mathbf{b} = \boldsymbol{\mu}_{-k} - \mathbf{a}\mu_k$,
- $\mathbf{c} = \mathbf{X}_k + \mathbf{X}_{-k}\mathbf{a}$,
- $\mathbf{d} = \mathbf{X}_{-k}\mathbf{b}_k$, and
- $\mathbf{s}_{-k|k}^2 = \text{dg}(\mathbf{X}_{-k}\boldsymbol{\Sigma}_{-k|k}\mathbf{X}_{-k}^T)$ where $\boldsymbol{\Sigma}_{-k|k} = \boldsymbol{\Sigma}_{-k,-k} - \Sigma_{k,k}\mathbf{a}\mathbf{a}^T$

need to be calculated for each $k = 1, \dots, p$ and each iteration. If calculated naively, this would cost $O(np^2)$ per index k and for each iteration, making it the most expensive cost of the algorithm for a total cost of $O(np^3)$. The vectors $\mathbf{a}$ and $\mathbf{b}$ can be calculated in $O(p)$ time and do not need to be considered further.

Suppose that $\boldsymbol{\eta}$, $\mathbf{M}$ and $\mathbf{s}^2$ have been pre-calculated. We first note that

$$\boldsymbol{\eta} = \mathbf{X}_k\mu_k + \mathbf{X}_{-k}\boldsymbol{\mu}_{-k}.$$

Hence, $\mathbf{X}_{-k}\boldsymbol{\mu}_{-k} = \boldsymbol{\eta} - \mathbf{X}_k\mu_k$. Next, we let

$$\begin{aligned}\mathbf{M} &= \begin{bmatrix} \mathbf{X}_k & \mathbf{X}_{-k} \end{bmatrix} \begin{bmatrix} \Sigma_{k,k} & \Sigma_{k,-k} \\ \Sigma_{-k,k} & \Sigma_{-k,-k} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{X}_k\Sigma_{k,k} + \mathbf{X}_{-k}\Sigma_{-k,k} & \mathbf{X}_k\Sigma_{k,-k} + \mathbf{X}_{-k}\Sigma_{-k,-k} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{M}_k & \mathbf{M}_{-k} \end{bmatrix}\end{aligned}$$

where $\mathbf{M}_k$ and $\mathbf{X}_k$ are the k -th columns of $\mathbf{M}$ and $\mathbf{X}$, respectively. Hence, $\mathbf{X}_{-k}\Sigma_{-k,k} = \mathbf{M}_k - \mathbf{X}_k\Sigma_{k,k}$. Next, we consider

$$\mathbf{M}\mathbf{X}^T = \mathbf{X}_k\Sigma_{k,k}\mathbf{X}_k^T + \mathbf{X}_{-k}\Sigma_{-k,k}\mathbf{X}_k^T + \mathbf{X}_k\Sigma_{k,-k}\mathbf{X}_{-k}^T + \mathbf{X}_{-k}\Sigma_{-k,-k}\mathbf{X}_{-k}^T$$

and hence,

$$\begin{aligned}\mathbf{s}^2 &= \Sigma_{k,k}\mathbf{X}_k^{\odot 2} + 2(\mathbf{X}_{-k}\Sigma_{-k,k}) \odot \mathbf{X}_k + \text{dg}(\mathbf{X}_{-k}\Sigma_{-k,-k}\mathbf{X}_{-k}^T) \\ &= \Sigma_{k,k}\mathbf{X}_k^{\odot 2} + 2(\mathbf{M}_k - \mathbf{X}_k\Sigma_{k,k}) \odot \mathbf{X}_k + \text{dg}(\mathbf{X}_{-k}\Sigma_{-k,-k}\mathbf{X}_{-k}^T) \\ &= 2(\mathbf{M}_k \odot \mathbf{X}_k) - \Sigma_{k,k}\mathbf{X}_k^{\odot 2} + \text{dg}(\mathbf{X}_{-k}\Sigma_{-k,-k}\mathbf{X}_{-k}^T)\end{aligned}$$

so that $\text{dg}(\mathbf{X}_{-k}\Sigma_{-k,-k}\mathbf{X}_{-k}^T) = \mathbf{s}^2 + \Sigma_{k,k}\mathbf{X}_k^{\odot 2} - 2(\mathbf{M}_k \odot \mathbf{X}_k)$. We can then calculate the vector $\mathbf{c}$ via

$$\begin{aligned}\mathbf{c} &= \mathbf{X}_k + \mathbf{X}_{-k}\Sigma_{-k,k}/\Sigma_{k,k} \\ &= \mathbf{X}_k + (\mathbf{M}_k - \mathbf{X}_k\Sigma_{k,k})/\Sigma_{k,k} \\ &= \mathbf{M}_k/\Sigma_{k,k}.\end{aligned}$$

Hence, when $\mathbf{M}$ is available the vector $\mathbf{c}$ can be calculated in $O(n)$ time rather than $O(np)$ time.

Next, we consider the vector $\mathbf{d}$. We have

$$\begin{aligned}\mathbf{d} &= \mathbf{X}_{-k}\boldsymbol{\mu}_{-k} - \mathbf{X}_{-k}\Sigma_{-k,k}\mu_k/\Sigma_{k,k} \\ &= \boldsymbol{\eta} - \mathbf{X}_k\mu_k - (\mathbf{M}_k - \mathbf{X}_k\Sigma_{k,k})\mu_k/\Sigma_{k,k} \\ &= \boldsymbol{\eta} - \mathbf{M}_k(\mu_k/\Sigma_{k,k}) \\ &= \boldsymbol{\eta} - \mathbf{c}\mu_k.\end{aligned}$$

Hence, when $\boldsymbol{\eta}$ and $\mathbf{M}$ are available we can calculate $\mathbf{d}$ in $O(n)$ time rather than $O(np)$ time.

Lastly, we have

$$\begin{aligned}
\mathbf{s}_{-k|k}^2 &= \text{dg}(\mathbf{X}_{-k} \boldsymbol{\Sigma}_{-k|k} \mathbf{X}_{-k}) \\
&= \text{dg}(\mathbf{X}_{-k} \boldsymbol{\Sigma}_{-k,-k} \mathbf{X}_{-k}) - \text{dg}(\mathbf{X}_{-k} \boldsymbol{\Sigma}_{-k,k} \boldsymbol{\Sigma}_{k,k}^{-1} \boldsymbol{\Sigma}_{k,-k} \mathbf{X}_{-k}) \\
&= \mathbf{s}^2 + \boldsymbol{\Sigma}_{k,k} \mathbf{X}_k^{\odot 2} - 2(\mathbf{M}_k \odot \mathbf{X}_k) - (\mathbf{M}_k - \mathbf{X}_k \boldsymbol{\Sigma}_{k,k})^{\odot 2} \boldsymbol{\Sigma}_{k,k}^{-1} \\
&= \mathbf{s}^2 - \mathbf{M}_k^{\odot 2} \boldsymbol{\Sigma}_{k,k}^{-1} \\
&= \mathbf{s}^2 - \mathbf{M}_k \odot \mathbf{c}
\end{aligned}$$

so that if $\mathbf{M}$ and $\mathbf{s}^2$ are available, then $\mathbf{s}_{-k|k}^2$ can be computed in $O(n)$ rather than $O(np^2)$ time.

Thus, using the precomputed quantities $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\mu}$, $\mathbf{M} = \mathbf{X}\boldsymbol{\Sigma}$, and $\mathbf{s}^2 = \text{dg}(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^T)$, the local GMP quantities are

$$\mathbf{a} = \frac{\boldsymbol{\Sigma}_{-k,k}}{\boldsymbol{\Sigma}_{k,k}}, \quad \mathbf{b} = \boldsymbol{\mu}_{-k} - \mathbf{a}\boldsymbol{\mu}_k, \quad \mathbf{c} = \frac{\mathbf{M}_k}{\boldsymbol{\Sigma}_{k,k}}, \quad \mathbf{d} = \boldsymbol{\eta} - \mathbf{c}\boldsymbol{\mu}_k,$$

and $\mathbf{s}_{-k|k}^2 = \mathbf{s}^2 - \mathbf{M}_k^{\odot 2} / \boldsymbol{\Sigma}_{k,k} = \mathbf{s}^2 - \mathbf{M}_k \odot \mathbf{c}$.

C.2 Updating joint Gaussian with new marginal distribution

Without loss of generality, suppose that we have partitioned the parameter vector $\boldsymbol{\theta}$ into $K = 2$ components $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with $q(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta} - \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ have been partitioned conformably with the partitioning of $\boldsymbol{\theta}$, i.e.,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Suppose that, through an MP update (or otherwise), $q(\boldsymbol{\theta}_1) = \phi(\boldsymbol{\theta}_1 - \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ is updated and replaced with $\tilde{q}(\boldsymbol{\theta}_1) = \phi(\boldsymbol{\theta}_1 - \tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$, with new (marginal) mean and covariance $\tilde{\boldsymbol{\mu}}_1$ and $\tilde{\boldsymbol{\Sigma}}_1$ respectively. Let $\tilde{q}(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta} - \tilde{\boldsymbol{\mu}}; \tilde{\boldsymbol{\Sigma}})$ denote the updated joint distribution obtained by replacing $q(\boldsymbol{\theta})$ with

$$\tilde{q}(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1) \tilde{q}(\boldsymbol{\theta}_1)$$

and the updated mean and covariance $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ is partitioned as

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \tilde{\boldsymbol{\mu}}_1 \\ \tilde{\boldsymbol{\mu}}_2 \end{bmatrix} \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{11} & \tilde{\boldsymbol{\Sigma}}_{12} \\ \tilde{\boldsymbol{\Sigma}}_{21} & \tilde{\boldsymbol{\Sigma}}_{22} \end{bmatrix}.$$

Note that $q(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1) = \phi(\boldsymbol{\theta}_2 - (\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\mu}_1)), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$. Then

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_2 &= \mathbb{E}_{\tilde{q}(\boldsymbol{\theta}_1)}[\mathbb{E}_q(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1)] \\ &= \mathbb{E}_{\tilde{q}(\boldsymbol{\theta}_1)}[\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\mu}_1)] \\ &= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\tilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1). \end{aligned}$$

Let d_1 and d_2 be the dimensions of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. This calculation can be performed in $O(d_1^3 + d_2d_1)$ time. Similarly,

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_{22} &= \mathbb{E}_{\tilde{q}(\boldsymbol{\theta}_1)}[\mathbb{V}_q(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)] + \mathbb{V}_{\tilde{q}(\boldsymbol{\theta}_1)}[\mathbb{E}_q(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)] \\ &= \mathbb{E}_{\tilde{q}(\boldsymbol{\theta}_1)}(\boldsymbol{\Sigma}_{2|1}) + \mathbb{V}_{\tilde{q}(\boldsymbol{\theta}_1)}[\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\mu}_1)] \\ &= \boldsymbol{\Sigma}_{2|1} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbb{V}_{\tilde{q}(\boldsymbol{\theta}_1)}[\boldsymbol{\theta}_1]\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ &= \boldsymbol{\Sigma}_{2|1} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\tilde{\boldsymbol{\Sigma}}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \end{aligned}$$

The update of $\tilde{\boldsymbol{\Sigma}}_{22}$ costs $O(d_1^3 + d_2^2d_1)$ time. Lastly,

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_{12} &= \mathbb{E}_{\tilde{q}(\boldsymbol{\theta}_1)}\left[\mathbb{E}_q\left\{(\boldsymbol{\theta}_1 - \mathbb{E}(\boldsymbol{\theta}_1))(\boldsymbol{\theta}_2 - \mathbb{E}(\boldsymbol{\theta}_2))^T \mid \boldsymbol{\theta}_1\right\}\right] \\ &= \mathbb{E}_{\tilde{q}(\boldsymbol{\theta}_1)}\left[(\boldsymbol{\theta}_1 - \tilde{\boldsymbol{\mu}}_1)\left(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\theta}_1 - \tilde{\boldsymbol{\mu}}_1)\right)^T\right] \\ &= \mathbb{E}_{\tilde{q}(\boldsymbol{\theta}_1)}\left[(\boldsymbol{\theta}_1 - \tilde{\boldsymbol{\mu}}_1)(\boldsymbol{\theta}_1 - \tilde{\boldsymbol{\mu}}_1)^T\right]\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ &= \tilde{\boldsymbol{\Sigma}}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \end{aligned}$$

The update of $\tilde{\boldsymbol{\Sigma}}_{12}$ costs $O(d_1^3 + d_1^2d_2)$ time.

C.3 Proof of Theorem 4

Proof 6 *Let*

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \mathbf{A} & \mathbf{C}^T \\ \mathbf{C} & \mathbf{D} \end{bmatrix}, \quad \text{where} \quad \mathbf{A} = \tilde{\boldsymbol{\Sigma}}_1.$$

Since $\mathbf{A}$ is positive definite, $\tilde{\Sigma}$ is positive definite if and only if its Schur complement $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{C}^T$ is positive definite. In the GMP update,

$$\mathbf{C} = \Sigma_{21}\Sigma_{11}^{-1}\tilde{\Sigma}_1 \quad \text{and} \quad \mathbf{D} = \Sigma_{2|1} + \Sigma_{21}\Sigma_{11}^{-1}\tilde{\Sigma}_1\Sigma_{11}^{-1}\Sigma_{12}.$$

Therefore, $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{C}^T = \Sigma_{2|1}$. Since Σ is positive definite, its Schur complement

$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

is positive definite. Hence, $\tilde{\Sigma}$ is positive definite.

C.4 Efficient updates

Here we concern ourselves with maintaining the quantities $\boldsymbol{\eta}$, $\mathbf{M}$ and $\mathbf{s}^2$ after the updates in Section C.2 have been performed. Using earlier notation

- $\tilde{\boldsymbol{\mu}}_{-k} = \mathbf{a}\tilde{\mu}_k + \mathbf{b}$
- $\tilde{\Sigma}_{k,-k} = \tilde{\Sigma}_k\mathbf{a}^T$, $\tilde{\Sigma}_{-k,k} = \tilde{\Sigma}_k\mathbf{a}$, and
- $\tilde{\Sigma}_{-k,-k} = \Sigma_{-k|k} + \tilde{\Sigma}_k\mathbf{a}\mathbf{a}^T = \Sigma_{-k,-k} + (\tilde{\Sigma}_k - \Sigma_{k,k})\mathbf{a}\mathbf{a}^T$.

Hence,

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma_{-k|k} & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} + \tilde{\Sigma}_k \begin{pmatrix} \mathbf{a} \\ 1 \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ 1 \end{pmatrix}^T.$$

Thus, for a scalar coordinate update, assuming $\Sigma_{-k|k}$ is available, the joint mean and covariance can be updated in $O(p^2)$ time.

Define $\tilde{\boldsymbol{\eta}} = \mathbf{X}\tilde{\boldsymbol{\mu}}$, $\tilde{\mathbf{M}} = \mathbf{X}\tilde{\Sigma}$, and $\tilde{\mathbf{s}}^2 = \text{dg}(\mathbf{X}\tilde{\Sigma}\mathbf{X}^T)$, where $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$ are the updated mean and

covariance matrix after the updates in Section C.2 have been applied. To update $\tilde{\boldsymbol{\eta}}$, we have

$$\begin{aligned}
\tilde{\boldsymbol{\eta}} &= \mathbf{X}_k \tilde{\boldsymbol{\mu}}_k + \mathbf{X}_{-k} \tilde{\boldsymbol{\mu}}_{-k} \\
&= \mathbf{X} \tilde{\boldsymbol{\mu}}_k + \mathbf{X}_{-k} [\boldsymbol{\mu}_{-k} + \boldsymbol{\Sigma}_{-k,k} \boldsymbol{\Sigma}_{k,k}^{-1} (\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)] \\
&= \boldsymbol{\eta} + \mathbf{X}_k (\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) + (\mathbf{M}_k - \mathbf{X}_k \boldsymbol{\Sigma}_{k,k}) \boldsymbol{\Sigma}_{k,k}^{-1} (\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) \\
&= \boldsymbol{\eta} + \mathbf{M}_k \boldsymbol{\Sigma}_{k,k}^{-1} (\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) \\
&= \boldsymbol{\eta} + (\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) \mathbf{c}
\end{aligned}$$

where the right-hand side costs $O(n)$ to calculate. Next we have

$$\widetilde{\mathbf{M}}_k = \mathbf{X}_k \widetilde{\boldsymbol{\Sigma}}_k + \mathbf{X}_{-k} \widetilde{\boldsymbol{\Sigma}}_{-k,k}, \quad \text{and} \quad \widetilde{\mathbf{M}}_{-k} = \mathbf{X}_k \widetilde{\boldsymbol{\Sigma}}_{k,-k} + \mathbf{X}_{-k} \widetilde{\boldsymbol{\Sigma}}_{-k,-k}.$$

Note that using the above expressions directly costs $O(np)$ to calculate $\widetilde{\mathbf{M}}_k$ and $O(np^2)$ to calculate $\widetilde{\mathbf{M}}_{-k}$. Dealing with $\widetilde{\mathbf{M}}_k$, we have

$$\begin{aligned}
\widetilde{\mathbf{M}}_k &= \mathbf{X}_k \widetilde{\boldsymbol{\Sigma}}_k + \mathbf{X}_{-k} \boldsymbol{\Sigma}_{-k,k} \boldsymbol{\Sigma}_{k,k}^{-1} \widetilde{\boldsymbol{\Sigma}}_k \\
&= \mathbf{X}_k \widetilde{\boldsymbol{\Sigma}}_k + (\mathbf{M}_k - \mathbf{X}_k \boldsymbol{\Sigma}_{k,k}) \boldsymbol{\Sigma}_{k,k}^{-1} \widetilde{\boldsymbol{\Sigma}}_k \\
&= \mathbf{M}_k \boldsymbol{\Sigma}_{k,k}^{-1} \widetilde{\boldsymbol{\Sigma}}_k \\
&= \mathbf{c} \widetilde{\boldsymbol{\Sigma}}_k.
\end{aligned}$$

This costs $O(n)$ time to calculate. Next, we consider $\widetilde{\mathbf{M}}_{-k}$ given by

$$\begin{aligned}
\widetilde{\mathbf{M}}_{-k} &= \mathbf{X}_k \widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_{k,k}^{-1} \boldsymbol{\Sigma}_{k,-k} + \mathbf{X}_{-k} [\boldsymbol{\Sigma}_{-k,-k} + \boldsymbol{\Sigma}_{-k,k} \boldsymbol{\Sigma}_{k,k}^{-1} (\widetilde{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_{k,k}) \boldsymbol{\Sigma}_{k,k}^{-1} \boldsymbol{\Sigma}_{k,-k}] \\
&= (\widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_{k,k}^{-1}) \mathbf{X}_k \boldsymbol{\Sigma}_{k,-k} + \mathbf{X}_{-k} \boldsymbol{\Sigma}_{-k,-k} + (\boldsymbol{\Sigma}_{k,k}^{-1} (\widetilde{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_{k,k}) \boldsymbol{\Sigma}_{k,k}^{-1}) \mathbf{X}_{-k} \boldsymbol{\Sigma}_{-k,k} \boldsymbol{\Sigma}_{k,-k} \\
&= (\widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_{k,k}^{-1}) \mathbf{X}_k \boldsymbol{\Sigma}_{k,-k} + \mathbf{M}_{-k} - \mathbf{X}_k \boldsymbol{\Sigma}_{k,-k} + (\boldsymbol{\Sigma}_{k,k}^{-1} (\widetilde{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_{k,k}) \boldsymbol{\Sigma}_{k,k}^{-1}) (\mathbf{M}_k - \mathbf{X}_k \boldsymbol{\Sigma}_{k,k}) \boldsymbol{\Sigma}_{k,-k} \\
&= \mathbf{M}_{-k} + (\boldsymbol{\Sigma}_{k,k}^{-1} \widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_{k,k}^{-1} - \boldsymbol{\Sigma}_{k,k}^{-1}) \mathbf{M}_k \boldsymbol{\Sigma}_{k,-k} \\
&= \mathbf{M}_{-k} + (\boldsymbol{\Sigma}_{k,k}^{-1} - \widetilde{\boldsymbol{\Sigma}}_k^{-1}) \widetilde{\mathbf{M}}_k \boldsymbol{\Sigma}_{k,-k}
\end{aligned}$$

which uses $\mathbf{X}_{-k} \boldsymbol{\Sigma}_{-k,-k} = \mathbf{M}_{-k} - \mathbf{X}_k \boldsymbol{\Sigma}_{k,-k}$ and $\mathbf{X}_{-k} \boldsymbol{\Sigma}_{-k,k} = \mathbf{M}_k - \mathbf{X}_k \boldsymbol{\Sigma}_{k,k}$. This term costs

$O(np)$ to compute. This is optimal, as $\widetilde{\mathbf{M}}_{-k}$ is an $n \times p$ matrix. Lastly, we have

$$\begin{aligned}
\widetilde{\mathbf{s}}^2 &= 2(\widetilde{\mathbf{M}}_k \odot \mathbf{X}_k) - \widetilde{\Sigma}_k \mathbf{X}_k^{\odot 2} + \text{dg}(\mathbf{X}_{-k} \widetilde{\Sigma}_{-k,-k} \mathbf{X}_{-k}^T) \\
&= 2(\widetilde{\mathbf{M}}_k \odot \mathbf{X}_k) - \widetilde{\Sigma}_k \mathbf{X}_k^{\odot 2} + \mathbf{s}_{-k|k}^2 + \widetilde{\Sigma}_k (\mathbf{X}_{-k} \Sigma_{-k,k} \Sigma_{k,k}^{-1})^{\odot 2} \\
&= 2(\Sigma_{k,k}^{-1} \widetilde{\Sigma}_k)(\mathbf{M}_k \odot \mathbf{X}_k) - \widetilde{\Sigma}_k \mathbf{X}_k^{\odot 2} + \mathbf{s}_{-k|k}^2 \\
&\quad + \Sigma_{k,k}^{-1} \widetilde{\Sigma}_k \Sigma_{k,k}^{-1} \mathbf{M}_k^{\odot 2} - 2\widetilde{\Sigma}_k \Sigma_{k,k}^{-1} (\mathbf{M}_k \odot \mathbf{X}_k) + \widetilde{\Sigma}_k \mathbf{X}_k^{\odot 2} \\
&= \mathbf{s}_{-k|k}^2 + (\Sigma_{k,k}^{-1} \widetilde{\Sigma}_k \Sigma_{k,k}^{-1}) \mathbf{M}_k^{\odot 2} \\
&= \mathbf{s}_{-k|k}^2 + \widetilde{\mathbf{M}}_k^{\odot 2} / \widetilde{\Sigma}_k
\end{aligned}$$

which can be calculated in $O(n)$ time.

Thus, we can update $\widetilde{\boldsymbol{\eta}}$, $\widetilde{\mathbf{M}}$ and $\widetilde{\mathbf{s}}^2$ via

$$\widetilde{\boldsymbol{\eta}} = \boldsymbol{\eta} + (\widetilde{\mu}_k - \mu_k) \mathbf{c}, \quad \widetilde{\mathbf{M}}_k = \mathbf{c} \widetilde{\Sigma}_k, \quad \widetilde{\mathbf{M}}_{-k} = \mathbf{M}_{-k} + \frac{\widetilde{\Sigma}_k - \Sigma_{k,k}}{\Sigma_{k,k}^2} \mathbf{M}_k \Sigma_{k,-k},$$

and $\widetilde{\mathbf{s}}^2 = \mathbf{s}_{-k|k}^2 + \widetilde{\mathbf{M}}_k^{\odot 2} / \widetilde{\Sigma}_k$.

Therefore, for each coordinate k , the local quantities required to form $r_k(\theta_k)$ can be obtained in $O(n)$ time once $\boldsymbol{\eta}$, $\mathbf{M}$, and $\mathbf{s}^2$ are maintained. Updating the Gaussian mean and covariance costs $O(p^2)$, while updating $\boldsymbol{\eta}$, $\mathbf{M}$, and $\mathbf{s}^2$ costs $O(np)$. Thus the deterministic algebraic cost of a full sweep over p coordinates is $O(np^2 + p^3)$, excluding the cost of normalizing and integrating the one-dimensional local densities.

D Additional details for linear models example

In this section, we provide detailed derivations for all of the material in Section 4.1. Throughout this section, $\text{IG}(a, b)$ denotes the inverse-gamma distribution with density proportional to $x^{-a-1} \exp(-b/x)$, so that

$$\mathbb{E}(X) = \frac{b}{a-1}, \quad \text{and} \quad \mathbb{V}(X) = \frac{b^2}{(a-1)^2(a-2)}.$$

D.1 Exact posterior distributions

The full conditional distributions are given by

$$\boldsymbol{\beta} | \mathbf{y}, \sigma^2 \sim N_p \left(u \hat{\boldsymbol{\beta}}, u \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \right) \quad \text{and} \quad \sigma^2 | \mathbf{y}, \boldsymbol{\beta} \sim \text{IG} \left(A + \frac{1}{2}(n+p), \mathcal{B}(\boldsymbol{\beta}) \right), \quad (30)$$

where $\mathcal{B}(\boldsymbol{\beta}) \equiv B + \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{1}{2g} \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, and $u = g/(1+g)$.

The exact posterior distributions for $\boldsymbol{\beta}$ and σ^2 are given by

$$\begin{aligned} \boldsymbol{\beta} | \mathbf{y} &\sim t \left(u \hat{\boldsymbol{\beta}}, \left(\frac{B + \frac{n}{2} \hat{\sigma}_u^2}{A + \frac{n}{2}} \right) u (\mathbf{X}^T \mathbf{X})^{-1}, 2A + n \right), \quad \text{and} \\ \sigma^2 | \mathbf{y} &\sim \text{IG} \left(A + \frac{n}{2}, B + \frac{n \hat{\sigma}_u^2}{2} \right), \end{aligned} \quad (31)$$

where

$$\hat{\sigma}_u^2 = \frac{1}{n} [\|\mathbf{y}\|^2 - u \mathbf{y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}]. \quad (32)$$

Note that as $g \rightarrow \infty$, $\hat{\sigma}_u^2$ converges to the maximum likelihood estimator for σ^2 .

D.2 MFVB for linear models

The MFVB approximation for the linear model corresponding to the factorization $q(\boldsymbol{\beta}, \sigma^2) = q(\boldsymbol{\beta})q(\sigma^2)$ has q -densities of the forms:

$$q(\boldsymbol{\beta}) = N_p(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \quad \text{and} \quad q(\sigma^2) = \text{IG}(\tilde{A}, \tilde{B}),$$

where the updates are given by equations (33) and (34) in Algorithm 1 which summarizes the MFVB method for the linear model. Note that the distributional form for $q(\beta)$ comes directly from (2), while the true posterior follows a t -distribution. Hence, for any factorized q -density MFVB cannot be exact for this model.

For the linear model described in Section 4.1 we use the q -density factorization $q(\beta, \sigma^2) = q(\beta)q(\sigma^2)$, the updates for $q(\beta)$ and $q(\sigma^2)$ are derived via

$$\begin{aligned}
q(\beta) &\propto \exp \left[\mathbb{E}_{q(\sigma^2)} \left\{ -\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\beta\|^2 - \frac{1}{2g\sigma^2} \beta^T \mathbf{X}^T \mathbf{X} \beta \right\} \right] \\
&\propto \exp \left[\mathbb{E}_{q(\sigma^2)} \left\{ \frac{\mathbf{y}^T \mathbf{X} \beta}{\sigma^2} - \frac{1+1/g}{2\sigma^2} \beta^T \mathbf{X}^T \mathbf{X} \beta \right\} \right] \\
&= N_p \left(\tilde{\boldsymbol{\mu}} \equiv u\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}} \equiv u(\tilde{B}/\tilde{A})(\mathbf{X}^T \mathbf{X})^{-1} \right) \quad \text{and} \\
q(\sigma^2) &\propto \exp \left[- \left(A + \frac{n+p}{2} + 1 \right) \log(\sigma^2) \right. \\
&\quad \left. - \mathbb{E}_{q(\beta)} \left\{ B + \frac{1}{2} \|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\mu}}\|^2 + \frac{1}{2g} \tilde{\boldsymbol{\mu}}^T \mathbf{X}^T \mathbf{X} \tilde{\boldsymbol{\mu}} + \frac{1}{2u} \text{tr}(\mathbf{X}^T \mathbf{X} \tilde{\boldsymbol{\Sigma}}) \right\} \sigma^{-2} \right] \\
&= \text{IG} \left(\tilde{A} \equiv A + \frac{n+p}{2}, \tilde{B} \equiv B + \frac{1}{2} \|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\mu}}\|^2 + \frac{1}{2g} \tilde{\boldsymbol{\mu}}^T \mathbf{X}^T \mathbf{X} \tilde{\boldsymbol{\mu}} + \frac{1}{2u} \text{tr}(\mathbf{X}^T \mathbf{X} \tilde{\boldsymbol{\Sigma}}) \right)
\end{aligned}$$

noting $1 + 1/g = 1/u$.

Algorithm 1 MFVB for the linear model

Require: $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $g > 0$, $A > 0$, and $B > 0$.

1: Calculate algorithm constants via (32).

2: Initialize variational parameter $\tilde{A} = A + \frac{1}{2}(n + p)$ and $\tilde{B} = B + \frac{1}{2}\|\mathbf{y}\|^2$

3: **repeat**

4: Update $q(\beta)$ parameters:

$$\tilde{\boldsymbol{\mu}} \leftarrow u\hat{\boldsymbol{\beta}}; \quad \tilde{\boldsymbol{\Sigma}} \leftarrow \frac{\tilde{B}}{\tilde{A}} u(\mathbf{X}^T \mathbf{X})^{-1} \quad (33)$$

5: Update $q(\sigma^2)$ parameters:

$$\tilde{A} \leftarrow A + \frac{n+p}{2}; \quad \tilde{B} \leftarrow B + \frac{\|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\mu}}\|^2}{2} + \frac{\tilde{\boldsymbol{\mu}}^T \mathbf{X}^T \mathbf{X} \tilde{\boldsymbol{\mu}}}{2g} + \frac{\text{tr}(\mathbf{X}^T \mathbf{X} \tilde{\boldsymbol{\Sigma}})}{2u} \quad (34)$$

6: **until** convergence criterion are met.

D.3 Derivations for MP - Approach 1

We consider the moment propagation approximation where $q(\beta) = N(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ and $q(\sigma^2) = \text{IG}(\tilde{A}, \tilde{B})$. These are precisely the same distributional forms as MFVB. However, the updates for the q -density parameters is different. For the update of $q(\beta)$ using the fact that

$\beta | \mathbf{y}, \sigma^2 \sim N(u\hat{\beta}, \sigma^2 u(\mathbf{X}^T \mathbf{X})^{-1})$, we equate

$$\mathbb{E}_r(\beta) = \mathbb{E}_q[\mathbb{E}(\beta | \mathbf{y}, \sigma^2)] = u\hat{\beta} \quad (35)$$

$$\mathbb{V}_r(\beta) = \mathbb{E}_q[\mathbb{V}(\beta | \mathbf{y}, \sigma^2)] + \mathbb{V}_q[\mathbb{E}(\beta | \mathbf{y}, \sigma^2)] = \frac{\tilde{B}}{\tilde{A} - 1} u(\mathbf{X}^T \mathbf{X})^{-1} \quad (36)$$

with $\mathbb{E}_q(\beta) = \tilde{\mu}$, and $\mathbb{V}_q(\beta) = \tilde{\Sigma}$. Hence, by matching (35) and (36) with $\mathbb{E}_q(\beta)$ and $\mathbb{V}_q(\beta)$, respectively, and solving for $\tilde{\mu}$ and $\tilde{\Sigma}$, we obtain the update

$$\tilde{\mu} \leftarrow u\hat{\beta} \quad \text{and} \quad \tilde{\Sigma} \leftarrow \frac{\tilde{B}}{\tilde{A} - 1} u(\mathbf{X}^T \mathbf{X})^{-1}.$$

Similarly, for the update of $q(\sigma^2)$ since $\sigma^2 | \mathbf{y}, \beta \sim \text{IG}(A + \frac{1}{2}(n + p), B(\beta))$ we have

$$\mathbb{E}_r(\sigma^2) = \mathbb{E}_q[\mathbb{E}(\sigma^2 | \mathbf{y}, \beta)] = \frac{\mathbb{E}_q[B(\beta)]}{A + \frac{n+p}{2} - 1}, \quad \text{and} \quad (37)$$

$$\begin{aligned} \mathbb{V}_r(\sigma^2) &= \mathbb{E}_q[\mathbb{V}(\sigma^2 | \mathbf{y}, \beta)] + \mathbb{V}_q[\mathbb{E}(\sigma^2 | \mathbf{y}, \beta)] \\ &= \mathbb{E}_q \left[\frac{B(\beta)^2}{(A + \frac{n+p}{2} - 1)^2 (A + \frac{n+p}{2} - 2)} \right] + \mathbb{V}_q \left[\frac{B(\beta)}{A + \frac{n+p}{2} - 1} \right] \\ &= \frac{\mathbb{E}_q[B(\beta)]^2}{(A + \frac{n+p}{2} - 1)^2 (A + \frac{n+p}{2} - 2)} \\ &\quad + \left[\frac{1}{(A + \frac{n+p}{2} - 1)^2 (A + \frac{n+p}{2} - 2)} + \frac{1}{(A + \frac{n+p}{2} - 1)^2} \right] \mathbb{V}_q[B(\beta)] \\ &= \frac{\mathbb{E}_q[B(\beta)]^2}{(A + \frac{n+p}{2} - 1)^2 (A + \frac{n+p}{2} - 2)} + \frac{\mathbb{V}_q[B(\beta)]}{(A + \frac{n+p}{2} - 1)(A + \frac{n+p}{2} - 2)}. \end{aligned} \quad (38)$$

Note that u and $B(\beta)$ are defined in Section D.1. Next,

$$\begin{aligned} \mathbb{E}_q[B(\beta)] &= B + \frac{1}{2} \|\mathbf{y} - \mathbf{X}\tilde{\mu}\|^2 + \frac{\tilde{\mu}^T \mathbf{X}^T \mathbf{X} \tilde{\mu}}{2g} + \frac{\text{tr}(\mathbf{X}^T \mathbf{X} \tilde{\Sigma})}{2u} \\ \mathbb{V}_q[B(\beta)] &= \mathbb{V}_q \left[B + \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \frac{\beta^T \mathbf{X}^T \mathbf{X} \beta}{2g} \right] \\ &= \mathbb{V}_q \left[\frac{1}{2} \beta^T (u^{-1} \mathbf{X}^T \mathbf{X}) \beta - \mathbf{y}^T \mathbf{X} \beta \right] \\ &= \frac{1}{4u^2} \mathbb{V}_q \left[(\beta - u\hat{\beta})^T (\mathbf{X}^T \mathbf{X}) (\beta - u\hat{\beta}) \right] \\ &= \frac{\text{tr}[(\mathbf{X}^T \mathbf{X} \tilde{\Sigma})^2]}{2u^2}. \end{aligned}$$

In the derivation of $\mathbb{V}_q[B(\beta)]$ above, the first line absorbs constants into the variance operator,

the second line completes the square in β , the third line takes constants outside the variance operator, and the last line follows from the variance formula for quadratic forms of multivariate normal variables, i.e., equation (19).

Hence, by matching (37) and (38) with $\mathbb{E}_q(\sigma^2)$ and $\mathbb{V}_q(\sigma^2)$, respectively, we obtain the following equations

$$\mathbb{E}_r(\sigma^2) = \frac{\tilde{B}}{\tilde{A} - 1} = \mathbb{E}_q(\sigma^2) \quad \text{and} \quad \mathbb{V}_r(\sigma^2) = \frac{\tilde{B}^2}{(\tilde{A} - 1)^2(\tilde{A} - 2)} = \mathbb{V}_q(\sigma^2).$$

Solving for $\tilde{A}$ and $\tilde{B}$ leads to the update

$$\tilde{A} \leftarrow \frac{[\mathbb{E}_r(\sigma^2)]^2}{\mathbb{V}_r(\sigma^2)} + 2; \quad \tilde{B} \leftarrow (\tilde{A} - 1) \mathbb{E}_r(\sigma^2). \quad (39)$$

The MP approximation can then be summarized via Algorithm 2.

D.4 Fixed-point properties of MP Approach 1

Result 1 *Let $a = A + n/2$ and $b = B + n\hat{\sigma}_u^2/2$. Assume $a > 2$. At any admissible fixed point of the moment conditions (40)–(46) of Algorithm 2, the propagated posterior mean and variance of σ^2 are*

$$\mathbb{E}_r^*(\sigma^2) = \frac{b}{a - 1} = \frac{B + n\hat{\sigma}_u^2/2}{A + n/2 - 1}, \quad \text{and} \quad \mathbb{V}_r^*(\sigma^2) = \frac{[\mathbb{E}_r^*(\sigma^2)]^2}{A + (n + p)/2 - 2} \left[1 + \frac{p/2}{A + (n + p)/2 - 1} \right]$$

respectively. Furthermore, $\tilde{\mu}^ = u\hat{\beta}$, and $\tilde{\Sigma}^* = \mathbb{E}_r^*(\sigma^2) u(\mathbf{X}^T \mathbf{X})^{-1}$. Hence the MP approximation gives the exact posterior means of β and σ^2 , and the exact posterior covariance matrix of β . However, the propagated variance $\mathbb{V}_r^*(\sigma^2)$ is smaller than the exact posterior variance $\mathbb{V}(\sigma^2|\mathbf{y}) = [\mathbb{E}(\sigma^2|\mathbf{y})]^2/(A + n/2 - 2)$. For fixed p , the ratio $\mathbb{V}_r^*(\sigma^2)/\mathbb{V}(\sigma^2|\mathbf{y})$ converges to 1 as $n \rightarrow \infty$. Thus, Algorithm 2 recovers the exact first posterior moments and the exact covariance of β , but it still underestimates the posterior variance of σ^2 by a term that vanishes asymptotically for fixed p .*

Algorithm 2 MP for the linear model: Approach 1

Require: $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $g > 0$, $A > 0$, and $B > 0$.

- 1: Calculate algorithm constants via (32).
- 2: Initialize variational parameters: $\tilde{A} = A + \frac{1}{2}(n + p)$ and $\tilde{B} = B + \frac{n}{2}\hat{\sigma}_u^2$.
- 3: **repeat**
- 4: Approximate posterior moments of β assuming $q(\sigma^2) = \text{IG}(\tilde{A}, \tilde{B})$, where $u = \frac{g}{1+g}$:

$$\mathbb{E}_r(\beta) \leftarrow u\hat{\beta}; \quad \mathbb{V}_r(\beta) \leftarrow \frac{\tilde{B}}{\tilde{A} - 1}u(\mathbf{X}^T\mathbf{X})^{-1} \quad (40)$$

- 5: Update $q(\beta)$ parameters via moment matching

$$\tilde{\mu} \leftarrow \mathbb{E}_r(\beta); \quad \tilde{\Sigma} \leftarrow \mathbb{V}_r(\beta) \quad (41)$$

- 6: Approximate posterior moments of σ^2 assuming $q(\beta) = N(\tilde{\mu}, \tilde{\Sigma})$:

$$\mathbb{E}_q[\mathcal{B}(\beta)] \leftarrow B + \frac{1}{2}\|\mathbf{y} - \mathbf{X}\tilde{\mu}\|^2 + \frac{1}{2g}\tilde{\mu}^T\mathbf{X}^T\mathbf{X}\tilde{\mu} + \frac{1}{2u}\text{tr}(\mathbf{X}^T\mathbf{X}\tilde{\Sigma}) \quad (42)$$

$$\mathbb{V}_q[\mathcal{B}(\beta)] \leftarrow \frac{\text{tr}[(\mathbf{X}^T\mathbf{X}\tilde{\Sigma})^2]}{2u^2} \quad (43)$$

$$\mathbb{E}_r(\sigma^2) \leftarrow \frac{\mathbb{E}_q[\mathcal{B}(\beta)]}{A + \frac{n+p}{2} - 1} \quad (44)$$

$$\mathbb{V}_r(\sigma^2) \leftarrow \frac{\mathbb{E}_q[\mathcal{B}(\beta)]^2}{(A + \frac{n+p}{2} - 1)^2(A + \frac{n+p}{2} - 2)} + \frac{\mathbb{V}_q[\mathcal{B}(\beta)]}{(A + \frac{n+p}{2} - 1)(A + \frac{n+p}{2} - 2)} \quad (45)$$

- 7: Update $q(\sigma^2)$ parameters via moment matching

$$\tilde{A} \leftarrow \frac{[\mathbb{E}_r(\sigma^2)]^2}{\mathbb{V}_r(\sigma^2)} + 2; \quad \tilde{B} \leftarrow \mathbb{E}_r(\sigma^2)(\tilde{A} - 1); \quad (46)$$

- 8: **until** convergence criteria are met.
-

Proof: At a fixed point of Algorithm 2, the left-hand sides of the assignments are equal to the corresponding right-hand sides of equations (40)–(46). This is equivalent to the following

system of equations where $\tilde{\boldsymbol{\mu}} = u\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\Sigma}} = \frac{\tilde{B}}{\tilde{A}-1}u(\mathbf{X}^T\mathbf{X})^{-1}$:

$$\begin{aligned}\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})] &= B + \frac{1}{2}\|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\mu}}\|^2 + \frac{1}{2g}\tilde{\boldsymbol{\mu}}^T\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\mu}} + \frac{1}{2u}\text{tr}(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\Sigma}}), \\ \mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})] &= \frac{1}{2u^2}\text{tr}\left[(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\Sigma}})^2\right], \\ \mathbb{E}_r(\sigma^2) &= \frac{\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})]}{A + \frac{n+p}{2} - 1}, \\ \mathbb{V}_r(\sigma^2) &= \frac{\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})]^2}{(A + \frac{n+p}{2} - 1)^2(A + \frac{n+p}{2} - 2)} + \frac{\mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})]}{(A + \frac{n+p}{2} - 1)(A + \frac{n+p}{2} - 2)}, \\ \tilde{A} &= \frac{[\mathbb{E}_r(\sigma^2)]^2}{\mathbb{V}_r(\sigma^2)} + 2, \quad \tilde{B} = (\tilde{A} - 1)\mathbb{E}_r(\sigma^2).\end{aligned}$$

Substituting $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ into the expressions for $\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})]$ and $\mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})]$, we obtain

$$\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})] = B + \frac{1}{2}n\hat{\sigma}_u^2 + \frac{1}{2}p\mathbb{E}_r(\sigma^2) \quad \text{and} \quad \mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})] = \frac{1}{2}p[\mathbb{E}_r(\sigma^2)]^2$$

which follows from the facts that

$$\begin{aligned}\|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\mu}}\|^2 + g^{-1}\tilde{\boldsymbol{\mu}}^T\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\mu}} &= n\hat{\sigma}_u^2 \\ u^{-1}\text{tr}\left[(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\Sigma}})\right] &= \left(\frac{\tilde{B}}{\tilde{A}-1}\right)p = p\mathbb{E}_r(\sigma^2) \\ u^{-2}\text{tr}\left[(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\Sigma}})^2\right] &= \left(\frac{\tilde{B}}{\tilde{A}-1}\right)^2 p = p[\mathbb{E}_r(\sigma^2)]^2 \\ \left(\frac{\tilde{B}}{\tilde{A}-1}\right) &= \mathbb{E}_r(\sigma^2).\end{aligned}$$

Substituting the above expressions for $\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})]$ and $\mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})]$ into the equations for $\mathbb{E}_r(\sigma^2)$ and $\mathbb{V}_r(\sigma^2)$ we obtain

$$\begin{aligned}\mathbb{E}_r(\sigma^2) &= \frac{B + \frac{n\hat{\sigma}_u^2}{2} + \frac{1}{2}p\mathbb{E}_r(\sigma^2)}{A + \frac{n+p}{2} - 1} \\ \mathbb{V}_r(\sigma^2) &= \frac{\left[B + \frac{n\hat{\sigma}_u^2}{2} + \frac{1}{2}p\mathbb{E}_r(\sigma^2)\right]^2}{(A + \frac{n+p}{2} - 1)^2(A + \frac{n+p}{2} - 2)} + \frac{\frac{p}{2}[\mathbb{E}_r(\sigma^2)]^2}{(A + \frac{n+p}{2} - 1)(A + \frac{n+p}{2} - 2)}.\end{aligned}\tag{47}$$

Solving (47) for $\mathbb{E}_r(\sigma^2)$ we obtain

$$\mathbb{E}_r(\sigma^2) = \frac{B + \frac{n\hat{\sigma}_u^2}{2}}{A + \frac{n}{2} - 1},$$

which is the true posterior mean of σ^2 . Furthermore, since the expression for $\mathbb{E}_r(\sigma^2)$ is exact, the expression for $\mathbb{E}_r(\beta)$ is also exact. Substituting $\mathbb{E}_r(\sigma^2)$ into the expression for $\mathbb{V}_r(\sigma^2)$, after simplification, we obtain

$$\mathbb{V}_r(\sigma^2) = \frac{1}{A + \frac{n+p}{2} - 2} \left(1 + \frac{\frac{p}{2}}{A + \frac{n+p}{2} - 1} \right) \left(\frac{B + \frac{n\hat{\sigma}_u^2}{2}}{A + \frac{n}{2} - 1} \right)^2.$$

From this, we can obtain expressions for $\tilde{A}$ and $\tilde{B}$ at convergence of Algorithm 2. However, the expressions $\tilde{A}$ and $\tilde{B}$ do not seem to lend themselves to additional insights.

The exact posterior variance is

$$\mathbb{V}(\sigma^2 | \mathbf{y}) = \frac{1}{A + \frac{n}{2} - 2} \left(\frac{B + \frac{n\hat{\sigma}_u^2}{2}}{A + \frac{n}{2} - 1} \right)^2.$$

To see the underestimation explicitly, write $t = p/2$. Then

$$\frac{\mathbb{V}_r^*(\sigma^2)}{\mathbb{V}(\sigma^2 | \mathbf{y})} = \frac{(a-2)(a+2t-1)}{(a+t-1)(a+t-2)}.$$

Since

$$(a+t-1)(a+t-2) - (a+2t-1)(a-2) = t(t+1) > 0,$$

this ratio is less than one for $p > 0$, and tends to one as $a \rightarrow \infty$ with p fixed.

Thus, Approach 1 corrects the posterior mean of σ^2 and the posterior mean and variance of β , but it does not recover the exact marginal variance of σ^2 . Approach 2 resolves this by using a multivariate t approximation for $q(\beta)$, matching the heavy-tailed exact marginal form.

D.5 Exact recovery for MP Approach 2

Now we have $q(\beta) = t_p(\tilde{\mu}, \tilde{\Sigma}, \tilde{\nu})$ and $q(\sigma^2) = \text{IG}(\tilde{A}, \tilde{B})$. To derive the update for β we first note that $\beta | \mathbf{y}, \sigma^2 \sim N(u\hat{\beta}, \sigma^2 u(\mathbf{X}^T \mathbf{X})^{-1})$. The mean and covariance of $r(\beta)$ are $\mathbb{E}_r(\beta) = u\hat{\beta}$, and $\mathbb{V}_r(\beta) = (\tilde{B}/(\tilde{A}-1))u(\mathbf{X}^T \mathbf{X})^{-1}$ as before. The corresponding mean and covariance of $q(\beta)$ are $\mathbb{E}_q(\beta) = \tilde{\mu}$ and $\mathbb{V}(\beta) = (\tilde{\nu}/(\tilde{\nu}-2))\tilde{\Sigma}$. Hence,

$$\tilde{\mu} = u\hat{\beta} \quad \text{and} \quad \tilde{\Sigma} = \left(\frac{\tilde{\nu}-2}{\tilde{\nu}} \right) \left(\frac{\tilde{B}}{\tilde{A}-1} \right) u(\mathbf{X}^T \mathbf{X})^{-1}.$$

Let $Q(\beta) = (\beta - \tilde{\mu})^T \tilde{\Sigma}^{-1} (\beta - \tilde{\mu})$. Then (20) implies

$$\mathbb{E}_q[Q(\beta)^2] = \frac{\tilde{\nu}^2(2p+p^2)}{(\tilde{\nu}-2)(\tilde{\nu}-4)}.$$

The fourth moment exists provided $\tilde{\nu} > 4$. We want to match this with $\mathbb{E}_r[Q(\beta)^2]$ where

$$Q(\beta) = \left(\frac{\tilde{\nu}}{\tilde{\nu}-2} \right) \left(\frac{\tilde{B}}{\tilde{A}-1} \right) (\beta - u\hat{\beta})^T \left(u(\mathbf{X}^T \mathbf{X})^{-1} \right)^{-1} (\beta - u\hat{\beta}).$$

Then using (19) where we replace Σ with $\sigma^2 u(\mathbf{X}^T \mathbf{X})^{-1}$, μ with $\mathbf{0}$, and $\mathbf{A} = [u(\mathbf{X}^T \mathbf{X})^{-1}]^{-1}$ we obtain

$$\begin{aligned} \mathbb{E}_r[Q(\beta)^2] &= \mathbb{E}_q[Q(\beta)^2 | \mathcal{D}, \sigma^2] \\ &= \left[\left(\frac{\tilde{\nu}}{\tilde{\nu}-2} \right) \left(\frac{\tilde{A}-1}{\tilde{B}} \right) \right]^2 (2p+p^2) \mathbb{E}_q(\sigma^4) \\ &= \left[\left(\frac{\tilde{\nu}}{\tilde{\nu}-2} \right) \left(\frac{\tilde{A}-1}{\tilde{B}} \right) \right]^2 (2p+p^2) \frac{\tilde{B}^2}{(\tilde{A}-1)(\tilde{A}-2)} \\ &= \left(\frac{\tilde{\nu}^2}{(\tilde{\nu}-2)^2} \right) \left(\frac{\tilde{A}-1}{\tilde{A}-2} \right) (2p+p^2). \end{aligned}$$

Equating $\mathbb{E}_r[Q(\beta)^2]$ with $\mathbb{E}_q[Q(\beta)^2]$, we have

$$\frac{\tilde{\nu}^2(2p+p^2)}{(\tilde{\nu}-2)(\tilde{\nu}-4)} = \left(\frac{\tilde{\nu}^2}{(\tilde{\nu}-2)^2} \right) \left(\frac{\tilde{A}-1}{\tilde{A}-2} \right) (2p+p^2)$$

which, after cancellation and rearranging, yields

$$\frac{\tilde{\nu} - 2}{\tilde{\nu} - 4} = \frac{\tilde{A} - 1}{\tilde{A} - 2}.$$

Solving for $\tilde{\nu}$, we obtain $\tilde{\nu} = 2\tilde{A}$. Hence, the update for $q(\beta)$ is given by

$$\tilde{\mu} \leftarrow u\hat{\beta}; \quad \tilde{\Sigma} \leftarrow \frac{\tilde{B}}{\tilde{A}}u(\mathbf{X}^T\mathbf{X})^{-1}; \quad \text{and} \quad \tilde{\nu} \leftarrow 2\tilde{A}.$$

Similarly, for the update of $q(\sigma^2)$ since $\sigma^2 | \mathbf{y}, \beta \sim \text{IG}(A + \frac{1}{2}(n+p), \mathcal{B}(\beta))$, we have

$$\begin{aligned} \mathbb{E}_r(\sigma^2) &= \mathbb{E}_q[\mathbb{E}(\sigma^2 | \mathbf{y}, \beta)] = \frac{\mathbb{E}_q[\mathcal{B}(\beta)]}{A + \frac{n+p}{2} - 1} \\ \mathbb{V}_r(\sigma^2) &= \frac{\mathbb{E}_q[\mathcal{B}(\beta)]^2}{(A + \frac{n+p}{2} - 1)^2(A + \frac{n+p}{2} - 2)} + \frac{\mathbb{V}_q[\mathcal{B}(\beta)]}{(A + \frac{n+p}{2} - 1)(A + \frac{n+p}{2} - 2)}. \end{aligned}$$

Next, using (21) we have

$$\begin{aligned} \mathbb{E}_q[\mathcal{B}(\beta)] &= B + \frac{1}{2}\|\mathbf{y} - \mathbf{X}\tilde{\mu}\|^2 + \frac{\tilde{\mu}^T \mathbf{X}^T \mathbf{X} \tilde{\mu}}{2g} + \frac{\tilde{\nu}}{\tilde{\nu} - 2} \frac{\text{tr}(\mathbf{X}^T \mathbf{X} \tilde{\Sigma})}{2u} \\ \mathbb{V}_q[\mathcal{B}(\beta)] &= \frac{1}{2u^2} \left[\frac{\tilde{\nu}^2 \text{tr}\{(\mathbf{X}^T \mathbf{X} \tilde{\Sigma})^2\}}{(\tilde{\nu} - 2)(\tilde{\nu} - 4)} + \frac{\tilde{\nu}^2 \{\text{tr}(\mathbf{X}^T \mathbf{X} \tilde{\Sigma})\}^2}{(\tilde{\nu} - 2)^2(\tilde{\nu} - 4)} \right]. \end{aligned}$$

We can perform the update for $q(\sigma^2)$ via moment matching with the inverse-gamma distribution via Equation (39).

Algorithm 3 MP for the linear model: Approach 2

Require: $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $g > 0$, $A > 0$, and $B > 0$.

1: Calculate algorithm constants via (32).

2: Initialize variational parameters: $\tilde{A} \leftarrow A + \frac{1}{2}(n+p)$ and $\tilde{B} \leftarrow B + \frac{n}{2}\hat{\sigma}_u^2$.

3: **repeat**

4: Update $q(\boldsymbol{\beta}) = t(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \tilde{\nu})$ via moment matching using approximate posterior means and covariances of $\boldsymbol{\beta}$, and moments of $Q(\boldsymbol{\beta})^2$ assuming $q(\sigma^2) = \text{IG}(\tilde{A}, \tilde{B})$:

$$\tilde{\boldsymbol{\mu}} \leftarrow u\hat{\boldsymbol{\beta}}; \quad \tilde{\boldsymbol{\Sigma}} \leftarrow \frac{\tilde{B}}{\tilde{A}}u(\mathbf{X}^T\mathbf{X})^{-1}; \quad \tilde{\nu} \leftarrow 2\tilde{A} \quad (48)$$

5: Approximate posterior moments of σ^2 assuming $q(\boldsymbol{\beta}) = t(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}, \tilde{\nu})$:

$$\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})] \leftarrow B + \frac{1}{2}\|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\mu}}\|^2 + \frac{1}{2g}\tilde{\boldsymbol{\mu}}^T\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\mu}} + \frac{\tilde{\nu}\text{tr}(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\Sigma}})}{2u(\tilde{\nu}-2)} \quad (49)$$

$$\mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})] \leftarrow \frac{1}{2u^2} \left[\frac{\tilde{\nu}^2\text{tr}\{(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\Sigma}})^2\}}{(\tilde{\nu}-2)(\tilde{\nu}-4)} + \frac{\tilde{\nu}^2\text{tr}\{(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\Sigma}})\}^2}{(\tilde{\nu}-2)^2(\tilde{\nu}-4)} \right] \quad (50)$$

$$\mathbb{E}_r(\sigma^2) \leftarrow \frac{\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})]}{A + \frac{n+p}{2} - 1}; \quad (51)$$

$$\mathbb{V}_r(\sigma^2) \leftarrow \frac{\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})]^2}{(A + \frac{n+p}{2} - 1)^2(A + \frac{n+p}{2} - 2)} + \frac{\mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})]}{(A + \frac{n+p}{2} - 1)(A + \frac{n+p}{2} - 2)} \quad (52)$$

6: Update $q(\sigma^2)$ parameters via moment matching:

$$\tilde{A} \leftarrow \frac{[\mathbb{E}_r(\sigma^2)]^2}{\mathbb{V}_r(\sigma^2)} + 2; \quad \tilde{B} \leftarrow \mathbb{E}_r(\sigma^2)(\tilde{A} - 1); \quad (53)$$

7: **until** convergence criteria are met.

D.6 MP Approach 2 Theory

Result 2 Suppose $\tilde{A} > 2$ and $\tilde{B} > 0$. The consistency conditions (48)–(53) of Algorithm 3 have a unique admissible fixed point, given by

$$q^*(\boldsymbol{\beta}) = t_p(\tilde{\boldsymbol{\mu}}^*, \tilde{\boldsymbol{\Sigma}}^*, \tilde{\nu}^*), \quad q^*(\sigma^2) = \text{IG}(\tilde{A}^*, \tilde{B}^*),$$

where

$$\begin{aligned} \tilde{A}^* &= A + \frac{n}{2}, & \tilde{B}^* &= B + \frac{n}{2}\hat{\sigma}_u^2, \\ \tilde{\boldsymbol{\mu}}^* &= u\hat{\boldsymbol{\beta}}, & \tilde{\boldsymbol{\Sigma}}^* &= \left(\frac{B + n\hat{\sigma}_u^2/2}{A + n/2} \right) u(\mathbf{X}^T\mathbf{X})^{-1}, & \tilde{\nu}^* &= 2A + n. \end{aligned}$$

These are the exact marginal posterior distributions for $\boldsymbol{\beta}$ and σ^2 .

Proof: At a fixed point of Algorithm 3, the left-hand sides of the assignments are equal to the corresponding right-hand sides of equations (48)–(53). This is equivalent to solving the following system of equations:

$$\tilde{\boldsymbol{\mu}} = u\hat{\boldsymbol{\beta}} \quad (54)$$

$$\tilde{\boldsymbol{\Sigma}} = \frac{\tilde{B}}{\tilde{A}}u(\mathbf{X}^T\mathbf{X})^{-1} \quad (55)$$

$$\tilde{\nu} = 2\tilde{A} \quad (56)$$

$$\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})] = B + \frac{1}{2}\|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\mu}}\|^2 + \frac{\tilde{\boldsymbol{\mu}}^T\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\mu}}}{2g} + \frac{\tilde{\nu}}{\tilde{\nu} - 2} \frac{\text{tr}(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\Sigma}})}{2u} \quad (57)$$

$$\mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})] = \frac{1}{2u^2} \left[\frac{\tilde{\nu}^2 \text{tr}[(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\Sigma}})^2]}{(\tilde{\nu} - 2)(\tilde{\nu} - 4)} + \frac{\tilde{\nu}^2 [\text{tr}(\mathbf{X}^T\mathbf{X}\tilde{\boldsymbol{\Sigma}})]^2}{(\tilde{\nu} - 2)^2(\tilde{\nu} - 4)} \right] \quad (58)$$

$$\mathbb{E}_r(\sigma^2) = \frac{\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})]}{A + \frac{n+p}{2} - 1} \quad (59)$$

$$\mathbb{V}_r(\sigma^2) = \frac{\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})]^2}{(A + \frac{n+p}{2} - 1)^2(A + \frac{n+p}{2} - 2)} + \frac{\mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})]}{(A + \frac{n+p}{2} - 1)(A + \frac{n+p}{2} - 2)} \quad (60)$$

$$\tilde{A} = \frac{[\mathbb{E}_r(\sigma^2)]^2}{\mathbb{V}_r(\sigma^2)} + 2 \quad (61)$$

$$\tilde{B} = (\tilde{A} - 1)\mathbb{E}_r(\sigma^2) \quad (62)$$

The moment-matching equations for $\boldsymbol{\beta}$, together with $\tilde{\nu} = 2\tilde{A}$, imply

$$\tilde{\boldsymbol{\Sigma}} = \frac{\tilde{\nu} - 2}{\tilde{\nu}}\mathbb{E}_r(\sigma^2)u(\mathbf{X}^T\mathbf{X})^{-1} \quad \text{and} \quad \frac{[\mathbb{E}_r(\sigma^2)]^2}{\tilde{\nu} - 4} = \frac{\mathbb{V}_r(\sigma^2)}{2}.$$

Substituting the above expressions and $\tilde{\boldsymbol{\mu}} = u\hat{\boldsymbol{\beta}}$ into the equations for $\mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})]$ and $\mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})]$ we obtain

$$\begin{aligned} \mathbb{E}_q[\mathcal{B}(\boldsymbol{\beta})] &= B + \frac{n}{2}\hat{\sigma}_u^2 + \frac{p}{2}\mathbb{E}_r(\sigma^2) \quad \text{and} \\ \mathbb{V}_q[\mathcal{B}(\boldsymbol{\beta})] &= \frac{p}{2}[\mathbb{E}_r(\sigma^2)]^2 + \frac{p^2 + 2p}{2} \frac{[\mathbb{E}_r(\sigma^2)]^2}{\tilde{\nu} - 4} = \frac{p}{2}[\mathbb{E}_r(\sigma^2)]^2 + \frac{p^2 + 2p}{4}\mathbb{V}_r(\sigma^2). \end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}_r(\sigma^2) &= \frac{B + \frac{n}{2}\hat{\sigma}_u^2 + \frac{p}{2}\mathbb{E}_r(\sigma^2)}{A + \frac{n+p}{2} - 1} \\ \mathbb{V}_r(\sigma^2) &= \frac{[B + \frac{n}{2}\hat{\sigma}_u^2 + \frac{p}{2}\mathbb{E}_r(\sigma^2)]^2}{(A + \frac{n+p}{2} - 1)^2(A + \frac{n+p}{2} - 2)} + \frac{\frac{p}{2}[\mathbb{E}_r(\sigma^2)]^2}{(A + \frac{n+p}{2} - 1)(A + \frac{n+p}{2} - 2)} \\ &\quad + \frac{\frac{p^2+2p}{4}\mathbb{V}_r(\sigma^2)}{(A + \frac{n+p}{2} - 1)(A + \frac{n+p}{2} - 2)}.\end{aligned}$$

Solving for the first equation gives

$$\mathbb{E}_r(\sigma^2) = \frac{B + \frac{n}{2}\hat{\sigma}_u^2}{A + \frac{n}{2} - 1}$$

which is the exact posterior mean for σ^2 .

To solve for $\mathbb{V}_r(\sigma^2)$ we next use $(A + \frac{n+p}{2} - 1)\mathbb{E}_r(\sigma^2) = B + \frac{n}{2}\hat{\sigma}_u^2 + \frac{p}{2}\mathbb{E}_r(\sigma^2)$ to obtain

$$\mathbb{V}_r(\sigma^2) = \frac{(A + \frac{n+p}{2} - 1 + \frac{p}{2})[\mathbb{E}_r(\sigma^2)]^2}{(A + \frac{n+p}{2} - 1)(A + \frac{n+p}{2} - 2)} + \frac{\frac{p^2+2p}{4}\mathbb{V}_r(\sigma^2)}{(A + \frac{n+p}{2} - 1)(A + \frac{n+p}{2} - 2)}.$$

Solving for $\mathbb{V}_r(\sigma^2)$ we have

$$\mathbb{V}_r(\sigma^2) = \frac{[\mathbb{E}_r(\sigma^2)]^2}{A + \frac{n}{2} - 2} = \frac{(B + \frac{n}{2}\hat{\sigma}_u^2)^2}{(A + \frac{n}{2} - 1)^2(A + \frac{n}{2} - 2)},$$

which is the exact posterior variance for σ^2 . Hence, the solution to the system of equations (54)-(62) is given by

$$\begin{aligned}\tilde{\boldsymbol{\mu}}^* &= u\hat{\boldsymbol{\beta}}, \quad \tilde{\boldsymbol{\Sigma}}^* = \left(\frac{B + \frac{n}{2}\hat{\sigma}_u^2}{A + \frac{n}{2}} \right) u(\mathbf{X}^T \mathbf{X})^{-1}, \quad \tilde{\nu}^* = 2A + n, \\ \tilde{A}^* &= A + \frac{n}{2} \quad \text{and} \quad \tilde{B}^* = B + \frac{n}{2}\hat{\sigma}_u^2,\end{aligned}$$

which correspond to the parameters of the exact posterior distribution where the starred values denote the values of the corresponding parameters of $q(\boldsymbol{\beta})$ and $q(\sigma^2)$ at convergence.

Uniqueness. Let $a = A + n/2$, $b = B + n\hat{\sigma}_u^2/2$, and $c = A + (n + p)/2 = a + p/2$. We show that the admissible fixed point satisfying $\tilde{A} > 2$ and $\tilde{B} > 0$ is unique. At any fixed point of Algorithm 3, Equation (48) determines values for $\tilde{\mu}$, $\tilde{\Sigma}$, and $\tilde{\nu}$. Therefore the only remaining free parameters are $\tilde{A}$ and $\tilde{B}$. Using these expressions in (49) gives

$$\mathbb{E}_q[\mathcal{B}(\beta)] = b + \frac{p\tilde{B}}{2(\tilde{A} - 1)}.$$

Indeed,

$$\frac{\tilde{\nu}}{\tilde{\nu} - 2} \frac{\text{tr}(\mathbf{X}^T \mathbf{X} \tilde{\Sigma})}{2u} = \frac{2\tilde{A}}{2\tilde{A} - 2} \frac{pu\tilde{B}/\tilde{A}}{2u} = \frac{p\tilde{B}}{2(\tilde{A} - 1)}.$$

The fixed-point condition for $\tilde{B}$ from (51) and (53) is

$$\tilde{B} = (\tilde{A} - 1) \frac{\mathbb{E}_q[\mathcal{B}(\beta)]}{c - 1} = \frac{\tilde{A} - 1}{c - 1} \left[b + \frac{p\tilde{B}}{2(\tilde{A} - 1)} \right].$$

Rearranging,

$$\tilde{B} \left[1 - \frac{p}{2(c - 1)} \right] = \frac{(\tilde{A} - 1)b}{c - 1}.$$

Since $c = a + p/2$, $1 - p/[2(c - 1)] = (a - 1)/(c - 1)$. Therefore, every admissible fixed point must satisfy

$$\tilde{B} = b \frac{\tilde{A} - 1}{a - 1}. \quad (63)$$

It remains to identify $\tilde{A}$. From (50), using $\tilde{\nu} = 2\tilde{A}$ and $\tilde{\Sigma} = (\tilde{B}/\tilde{A})u(\mathbf{X}^T \mathbf{X})^{-1}$, we obtain

$$\mathbb{V}_q[\mathcal{B}(\beta)] = \tilde{B}^2 \frac{p\{2(\tilde{A} - 1) + p\}}{4(\tilde{A} - 1)^2(\tilde{A} - 2)}.$$

Also, equations (51)–(53) imply

$$\mathbb{E}_r(\sigma^2) = \frac{\tilde{B}}{\tilde{A} - 1}, \quad \mathbb{V}_r(\sigma^2) = \frac{[\mathbb{E}_r(\sigma^2)]^2}{\tilde{A} - 2} = \frac{\tilde{B}^2}{(\tilde{A} - 1)^2(\tilde{A} - 2)}. \quad (64)$$

On the other hand, substituting (49) and (52) gives

$$\mathbb{V}_r(\sigma^2) = \frac{\tilde{B}^2}{(\tilde{A} - 1)^2(c - 2)} + \frac{\mathbb{V}_q[\mathcal{B}(\beta)]}{(c - 1)(c - 2)}.$$

Combining this expression with (64), and canceling the positive factor $\tilde{B}^2/(\tilde{A} - 1)^2$, yields

$$\frac{1}{\tilde{A} - 2} = \frac{1}{c - 2} + \frac{p[2(\tilde{A} - 1) + p]}{4(\tilde{A} - 2)(c - 1)(c - 2)}.$$

Multiplying by $(\tilde{A} - 2)(c - 2)$ gives the linear equation

$$\tilde{A} - c + \frac{p[2(\tilde{A} - 1) + p]}{4(c - 1)} = 0.$$

Equivalently, $[4(c - 1) + 2p]\tilde{A} = 4c(c - 1) - p(p - 2) = 4a(a + p - 1)$. Since $c = a + p/2$, the right-hand side simplifies to and the coefficient of $\tilde{A}$ is $4(c - 1) + 2p = 4(a + p - 1)$. Thus,

$$\tilde{A} = a = A + \frac{n}{2}.$$

Substituting this into (63) gives

$$\tilde{B} = b = B + \frac{n}{2}\hat{\sigma}_u^2.$$

Therefore $\tilde{A}$ and $\tilde{B}$ are uniquely determined. Equation (48) then uniquely determines $\tilde{\mu}$, $\tilde{\Sigma}$, and $\tilde{\nu}$. Hence the admissible fixed point is unique.

E MP Derivations for MVN model

In this appendix, we provide detailed derivations for all of the material in Section 4.2. The MP updates below involve fourth moments of a multivariate t distribution and second moments of inverse-Wishart random matrices. We therefore assume $\nu_n > p + 3$, so that the required moments exist at the exact posterior fixed point.

E.1 Exact posterior distributions for the MVN model

We next state the full conditional distributions for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which are useful for deriving both the MFVB and MP approximations:

$$\boldsymbol{\mu} \mid \mathbf{X}, \boldsymbol{\Sigma} \sim N_p(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}/\lambda_n) \quad \text{and} \quad \boldsymbol{\Sigma} \mid \mathbf{X}, \boldsymbol{\mu} \sim \text{IW}_p(\boldsymbol{\Psi}(\boldsymbol{\mu}), \nu_n + 1)$$

where $\boldsymbol{\Psi}(\boldsymbol{\mu}) = \boldsymbol{\Psi}_0 + \mathbf{S} + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})^T + \lambda_0 \boldsymbol{\mu} \boldsymbol{\mu}^T$,

$$\begin{aligned} \bar{\mathbf{x}} &= \mathbf{X}^T \mathbf{1}_n / n, \quad \lambda_n = \lambda_0 + n; \quad \nu_n = \nu_0 + n, \\ \boldsymbol{\mu}_n &= n\bar{\mathbf{x}}/\lambda_n, \quad \mathbf{S} = \mathbf{X}^T \mathbf{X} - n\bar{\mathbf{x}} \bar{\mathbf{x}}^T, \quad \text{and} \quad \boldsymbol{\Psi}_n = \boldsymbol{\Psi}_0 + \mathbf{S} + n\lambda_0/\lambda_n \bar{\mathbf{x}} \bar{\mathbf{x}}^T. \end{aligned} \tag{65}$$

For this example, we are able to calculate the exact posterior distribution and its corresponding moments. This can be used as a gold standard for assessing the quality of MFVB and MP algorithms. The marginal posterior distribution of $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma} \mid \mathbf{X} \sim \text{IW}_p(\boldsymbol{\Psi}_n, \nu_n)$.

The marginal posterior distribution for $\boldsymbol{\mu}$ can be found via

$$\boldsymbol{\mu} \mid \mathbf{X} \sim t_p\left(\boldsymbol{\mu}_n, \frac{\boldsymbol{\Psi}_n}{\lambda_n(\nu_n - p + 1)}, \nu_n - p + 1\right),$$

where we have used the usual parametrization of the multivariate t distribution whose density and various expectations are summarised in section A.2 for reference.

E.2 MFVB Derivations for the MVN model

For the MVN model in Section E.1 with q -density factorization $q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = q(\boldsymbol{\mu})q(\boldsymbol{\Sigma})$, the updates for $q(\boldsymbol{\mu})$ and $q(\boldsymbol{\Sigma})$ are derived via

$$\begin{aligned}
q(\boldsymbol{\mu}) &\propto \exp \left[-\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) - \frac{1}{2} \lambda_0 \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right] \\
&\propto \exp \left[-\frac{\lambda_n}{2} \boldsymbol{\mu}^T \mathbb{E}_q(\boldsymbol{\Sigma}^{-1}) \boldsymbol{\mu} + \boldsymbol{\mu}^T \mathbb{E}_q(\boldsymbol{\Sigma}^{-1}) n \bar{\mathbf{x}} \right] \\
&\propto \exp \left[-\frac{\lambda_n \tilde{d}}{2} \boldsymbol{\mu}^T \tilde{\boldsymbol{\Psi}}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^T \tilde{\boldsymbol{\Psi}}^{-1} \tilde{d} n \bar{\mathbf{x}} \right] \\
&\sim N_p(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \\
q(\boldsymbol{\Sigma}) &\propto \exp \left[-\frac{n+1+\nu_0+p+1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \text{tr} \left\{ \boldsymbol{\Sigma}^{-1} \mathbb{E}_q(\boldsymbol{\Psi}(\boldsymbol{\mu})) \right\} \right] \\
&\sim \text{IW}_p(\tilde{\boldsymbol{\Psi}}, \tilde{d}).
\end{aligned}$$

where the update for $q(\boldsymbol{\mu})$ is given by $\tilde{\boldsymbol{\mu}} \leftarrow \boldsymbol{\mu}_n$ and $\tilde{\boldsymbol{\Sigma}} \leftarrow \tilde{\boldsymbol{\Psi}}/(\lambda_n \tilde{d})$, and the update for $q(\boldsymbol{\Sigma})$ is given by $\tilde{\boldsymbol{\Psi}} \leftarrow \boldsymbol{\Psi}_n + \lambda_n \tilde{\boldsymbol{\Sigma}}$ and $\tilde{d} \leftarrow \nu_n + 1$. Algorithm 4 summarises the MFVB updates.

Algorithm 4 MFVB method for the MVN model

Require: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\Psi}_0 \in \mathbb{S}_+^p$, $\lambda_0 > 0$, $\nu_0 > p - 1$.

- 1: Calculate algorithm constants via (65).
- 2: Initialize: $\tilde{d} \leftarrow \nu_n + 1$ and $\tilde{\boldsymbol{\Psi}} \leftarrow \boldsymbol{\Psi}_n$.
- 3: **repeat**
- 4: Update $q(\boldsymbol{\mu})$:

$$\tilde{\boldsymbol{\mu}} \leftarrow \boldsymbol{\mu}_n; \quad \tilde{\boldsymbol{\Sigma}} \leftarrow \frac{\tilde{\boldsymbol{\Psi}}}{\lambda_n \tilde{d}}; \quad (66)$$

- 5: Update $q(\boldsymbol{\Sigma})$:

$$\tilde{d} \leftarrow \nu_n + 1; \quad \tilde{\boldsymbol{\Psi}} \leftarrow \boldsymbol{\Psi}_n + \lambda_n \tilde{\boldsymbol{\Sigma}}; \quad (67)$$

- 6: **until** until convergence criteria are met.
-

E.3 Derivations of MP update of $q(\boldsymbol{\mu})$ for the MVN model

Equating $\mathbb{E}_r(\boldsymbol{\mu})$ with $\mathbb{E}_q(\boldsymbol{\mu})$, and $\mathbb{V}_r(\boldsymbol{\mu})$ with $\mathbb{V}_q(\boldsymbol{\mu})$ we have

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_n \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}} = \frac{\tilde{\nu} - 2}{\tilde{\nu} \lambda_n (\Delta - 1)} \tilde{\boldsymbol{\Psi}} \quad (68)$$

where $\Delta = \tilde{d} - p$. Using (20) with $\boldsymbol{\mu} = \mathbf{0}$, $a = 1$, $\mathbf{A} = \mathbf{I}$, $\boldsymbol{\Sigma} = \tilde{\boldsymbol{\Sigma}}$ and $b = \tilde{\nu}$ we obtain

$$\mathbb{E}_q [\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^4] = \frac{\tilde{\nu}^2 [2 \operatorname{tr}(\tilde{\boldsymbol{\Sigma}}^2) + \operatorname{tr}(\tilde{\boldsymbol{\Sigma}})^2]}{(\tilde{\nu} - 2)(\tilde{\nu} - 4)} \quad (69)$$

and using (68) we have

$$\mathbb{E}_q [\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^4] = \frac{(\tilde{\nu} - 2) [2 \operatorname{tr}(\tilde{\boldsymbol{\Psi}}^2) + \operatorname{tr}(\tilde{\boldsymbol{\Psi}})^2]}{\lambda_n^2 (\Delta - 1)^2 (\tilde{\nu} - 4)}. \quad (70)$$

Using Corollary 3.1 (i) and (iv) from von Rosen (1988) summarized in Section A.3 we have

$$\mathbb{E} [\operatorname{tr}(\boldsymbol{\Sigma}^2)] = (c_1 + c_2) \operatorname{tr}(\tilde{\boldsymbol{\Psi}}^2) + c_2 \operatorname{tr}(\tilde{\boldsymbol{\Psi}})^2 \quad \text{and} \quad \mathbb{E} [\operatorname{tr}(\boldsymbol{\Sigma})^2] = c_1 \operatorname{tr}(\tilde{\boldsymbol{\Psi}})^2 + 2c_2 \operatorname{tr}(\tilde{\boldsymbol{\Psi}}^2)$$

where $c_2^{-1} = \Delta(\Delta - 1)(\Delta - 3)$, $c_1 = (\Delta - 2)c_2$ and provided $\Delta > 3$. Hence, we obtain

$$\begin{aligned} \mathbb{E}_r [\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^4] &= \mathbb{E}_q [\mathbb{E} \{ \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^4 \mid \mathbf{X}, \boldsymbol{\Sigma} \}] \\ &= \lambda_n^{-2} \mathbb{E}_q [\operatorname{tr}(\boldsymbol{\Sigma})^2 + 2 \operatorname{tr}(\boldsymbol{\Sigma}^2)] \\ &= \lambda_n^{-2} [c_1 \operatorname{tr}(\tilde{\boldsymbol{\Psi}})^2 + 2c_2 \operatorname{tr}(\tilde{\boldsymbol{\Psi}}^2)] + \lambda_n^{-2} [(2c_1 + 2c_2) \operatorname{tr}(\tilde{\boldsymbol{\Psi}}^2) + 2c_2 \operatorname{tr}(\tilde{\boldsymbol{\Psi}})^2] \\ &= \lambda_n^{-2} (c_1 + 2c_2) [2 \operatorname{tr}(\tilde{\boldsymbol{\Psi}}^2) + \operatorname{tr}(\tilde{\boldsymbol{\Psi}})^2] \\ &= \frac{2 \operatorname{tr}(\tilde{\boldsymbol{\Psi}}^2) + \operatorname{tr}(\tilde{\boldsymbol{\Psi}})^2}{\lambda_n^2 (\Delta - 1)(\Delta - 3)} \end{aligned} \quad (71)$$

noting that $c_1 + 2c_2 = (\Delta - 2)c_2 + 2c_2 = \Delta c_2 = 1/((\Delta - 1)(\Delta - 3))$.

Equating (70) with the final expression in (71), and substituting the expression for $\tilde{\boldsymbol{\Sigma}}$ in (68), we obtain

$$\frac{(\tilde{\nu} - 2) [2 \operatorname{tr}(\tilde{\boldsymbol{\Psi}}^2) + \operatorname{tr}(\tilde{\boldsymbol{\Psi}})^2]}{\lambda_n^2 (\Delta - 1)^2 (\tilde{\nu} - 4)} = \frac{2 \operatorname{tr}(\tilde{\boldsymbol{\Psi}}^2) + \operatorname{tr}(\tilde{\boldsymbol{\Psi}})^2}{\lambda_n^2 (\Delta - 1)(\Delta - 3)}$$

after solving for $\tilde{\nu}$ we find $\tilde{\nu} = \Delta + 1 = \tilde{d} - p + 1$.

E.4 MP posterior expectation of $q(\Sigma)$

Since, $\Sigma | \mathbf{X}, \boldsymbol{\mu} \sim \text{IW}_p(\Psi(\boldsymbol{\mu}), \nu_n + 1)$ the corrected posterior expectation of Σ is given by

$$\begin{aligned} \mathbb{E}_r(\Sigma) &= \mathbb{E}_q [\mathbb{E}(\Sigma | \mathbf{X}, \boldsymbol{\mu})] \\ &= \frac{1}{\nu_n + 1 - p - 1} \mathbb{E}_q [\Psi(\boldsymbol{\mu})] \\ &= \frac{1}{\nu_n - p} \left(\Psi_n + \frac{\tilde{\nu}}{\tilde{\nu} - 2} \lambda_n \tilde{\Sigma} \right) \\ &= \frac{\mathbf{A}}{\nu_n - p} \end{aligned}$$

where $\mathbb{E}_q [\Psi(\boldsymbol{\mu})] = \Psi_n + \frac{\lambda_n \tilde{\nu}}{\tilde{\nu} - 2} \tilde{\Sigma} \equiv \mathbf{A}$.

E.5 MP posterior variance of elements of Σ

In the following derivations the vector $\mathbf{e}_i$ is a vector of zeros, except for the i th element which is equal to 1. The matrix $\mathbf{E}_{ij}$ is a matrix whose elements are 0 except for the (i, j) th element which is equal to 1.

The MP variance of the diagonal elements of Σ are given by

$$\begin{aligned} \mathbb{V}_r(\Sigma_{ii}) &= \mathbb{E}_q [\mathbb{V}(\Sigma_{ii} | \mathbf{X}, \boldsymbol{\mu})] + \mathbb{V}_q [\mathbb{E}(\Sigma_{ii} | \mathbf{X}, \boldsymbol{\mu})] \\ &= \mathbb{E}_q \left[\frac{2\{\Psi(\boldsymbol{\mu})\}_{ii}^2}{(\nu_n - p)^2(\nu_n - p - 2)} \right] + \mathbb{V}_q \left[\frac{\{\Psi(\boldsymbol{\mu})\}_{ii}}{\nu_n - p} \right] \\ &= \frac{2(a_{ii}^2 + b_{ii})}{(\nu_n - p)^2(\nu_n - p - 2)} + \frac{b_{ii}}{(\nu_n - p)^2} \\ &= \frac{2a_{ii}^2 + (\nu_n - p)b_{ii}}{(\nu_n - p)^2(\nu_n - p - 2)} \end{aligned}$$

where $\{\Psi(\boldsymbol{\mu})\}_{ij} = \psi_{0,ij} + s_{ij} + n(\bar{x}_i - \mu_i)(\bar{x}_j - \mu_j) + \lambda_0 \mu_i \mu_j$,

$$a_{ij} \equiv \mathbb{E}_q [\{\Psi(\boldsymbol{\mu})\}_{ij}] \quad \text{and} \quad b_{ij} \equiv \mathbb{V}_q [\{\Psi(\boldsymbol{\mu})\}_{ij}].$$

Hence, we can write

$$\text{dg}[\mathbb{E}\mathbb{W}\mathbb{V}_r(\Sigma)] = \frac{2\text{dg}(\mathbf{A})^{\odot 2} + (\nu_n - p)\text{dg}(\mathbf{B})}{(\nu_n - p)^2(\nu_n - p - 2)}.$$

We have shown that $\mathbb{E}_q[\{\Psi(\boldsymbol{\mu})\}_{ij}] = a_{ij}$, but we have yet to derive an expression for b_{ij} .

Expanding and completing the square for $\Psi(\boldsymbol{\mu})$ we obtain

$$\begin{aligned}\Psi(\boldsymbol{\mu}) &= \Psi_0 + \mathbf{X}^T \mathbf{X} - n \bar{\mathbf{x}} \bar{\mathbf{x}}^T + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})^T + \lambda_0 \boldsymbol{\mu} \boldsymbol{\mu}^T \\ &= \Psi_n + \lambda_n (\boldsymbol{\mu} - \boldsymbol{\mu}_n)(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T.\end{aligned}$$

Next, we use (21) to calculate

$$\begin{aligned}\mathbb{V}_q[\{\Psi(\boldsymbol{\mu})\}_{ij}] &= \lambda_n^2 \mathbb{V}_q[(\mu_i - \tilde{\mu}_i)(\mu_j - \tilde{\mu}_j)] \\ &= \lambda_n^2 \mathbb{V}_q[(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) \mathbf{e}_i \mathbf{e}_j^T (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})] \\ &= \lambda_n^2 \mathbb{V}_q[(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) \mathbf{E}_{ij} (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})].\end{aligned}$$

The diagonal elements of $\mathbf{B}$ are given by

$$\begin{aligned}\mathbb{V}_q[\{\Psi(\boldsymbol{\mu})\}_{ii}] &= \lambda_n^2 \mathbb{V}_q[(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) \mathbf{E}_{ii} (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})] \\ &= \lambda_n^2 \left[\frac{2\tilde{\nu}^2 \text{tr}(\mathbf{E}_{ii} \tilde{\boldsymbol{\Sigma}} \mathbf{E}_{ii} \tilde{\boldsymbol{\Sigma}})}{(\tilde{\nu} - 2)(\tilde{\nu} - 4)} + \frac{2\tilde{\nu}^2 \text{tr}(\mathbf{E}_{ii} \tilde{\boldsymbol{\Sigma}})^2}{(\tilde{\nu} - 2)^2(\tilde{\nu} - 4)} \right] \\ &= \lambda_n^2 \left[\frac{2\tilde{\nu}^2 \tilde{\Sigma}_{ii}^2}{(\tilde{\nu} - 2)(\tilde{\nu} - 4)} + \frac{2\tilde{\nu}^2 \tilde{\Sigma}_{ii}^2}{(\tilde{\nu} - 2)^2(\tilde{\nu} - 4)} \right] \\ &= \frac{2\lambda_n^2 \tilde{\nu}^2 (\tilde{\nu} - 1) \tilde{\Sigma}_{ii}^2}{(\tilde{\nu} - 2)^2(\tilde{\nu} - 4)} \\ &\equiv b_{ii}.\end{aligned}$$

Hence,

$$\mathbf{b} \equiv \text{dg}[\mathbb{E} \mathbb{W} \mathbb{V}_q \{\Psi(\boldsymbol{\mu})\}] = \frac{2\lambda_n^2 \tilde{\nu}^2 (\tilde{\nu} - 1)}{(\tilde{\nu} - 2)^2(\tilde{\nu} - 4)} \text{dg}(\tilde{\boldsymbol{\Sigma}})^{\odot 2}.$$

Algorithm 5 summarizes the MP algorithm for the multivariate normal model.

Algorithm 5 Moment Propagation algorithm for the multivariate normal model

Require: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\Psi_0 \in \mathbb{S}_+^p$, $\lambda_0 > 0$, $\nu_0 > p - 1$.

- 1: Calculate algorithm constants via (65).
- 2: Initialize $q(\Sigma)$: $\tilde{d} \leftarrow \nu_n$, and $\tilde{\Psi} \leftarrow \Psi_n$.
- 3: **repeat**
- 4: Update $q(\mu) = t_p(\tilde{\mu}, \tilde{\Sigma}, \tilde{\nu})$ via moment matching using approximate posterior means and covariances of μ , and moments of $\|\mu - \tilde{\mu}\|^4$ assuming $q(\Sigma) = \text{IW}_p(\tilde{\Psi}, \tilde{d})$:

$$\tilde{\mu} \leftarrow \mu_n; \quad \tilde{\Sigma} \leftarrow \frac{\tilde{\Psi}}{\lambda_n(\tilde{d} - p + 1)}; \quad \tilde{\nu} \leftarrow \tilde{d} - p + 1; \quad (72)$$

- 5: Approximate posterior moments of Σ assuming $q(\mu) = t_p(\tilde{\mu}, \tilde{\Sigma}, \tilde{\nu})$:

$$\begin{aligned} \mathbb{E}_r(\Sigma) &\leftarrow \frac{\mathbf{A}}{\nu_n - p} \quad \text{and} \\ \text{dg}[\text{EWW}_r(\Sigma)] &\leftarrow \frac{2\text{dg}(\mathbf{A})^{\odot 2} + (\nu_n - p)\mathbf{b}}{(\nu_n - p)^2(\nu_n - p - 2)} \end{aligned} \quad (73)$$

where

$$\mathbf{A} = \Psi_n + \frac{\lambda_n \tilde{\nu}}{\tilde{\nu} - 2} \tilde{\Sigma}, \quad \text{and} \quad \mathbf{b} = \frac{2\lambda_n^2 \tilde{\nu}^2 (\tilde{\nu} - 1)}{(\tilde{\nu} - 2)^2 (\tilde{\nu} - 4)} \text{dg}(\tilde{\Sigma})^{\odot 2}. \quad (74)$$

- 6: Perform the MP update for $q(\Sigma)$ via moment matching:

$$\tilde{d} \leftarrow \frac{2\mathbf{1}^T [\text{dg}(\mathbb{E}_r(\Sigma))^{\odot 2}]}{\mathbf{1}_p^T \text{dg}[\text{EWW}_r(\Sigma)]} + p + 3; \quad \tilde{\Psi} \leftarrow (\tilde{d} - p - 1) \mathbb{E}_r(\Sigma); \quad (75)$$

- 7: **until** convergence criteria are met
-

E.6 Result 3

Result 3 Assume $\nu_n > p + 3$, so that the fourth moments required by Algorithm 5 exist. Within the admissible region $\tilde{d} > p + 3$, the moment equations (72)-(75) have a unique fixed point,

$$\tilde{\Psi}^* = \Psi_n, \quad \tilde{d}^* = \nu_n, \quad \tilde{\mu}^* = \mu_n, \quad \tilde{\Sigma}^* = \frac{\Psi_n}{\lambda_n(\nu_n - p + 1)}, \quad \tilde{\nu}^* = \nu_n - p + 1.$$

This fixed point corresponds to the exact marginal posterior distributions of μ and Σ . The algebraic fixed-point equation also has a boundary root $\tilde{d} = p + 3$, equivalently $\tilde{\nu} = 4$, but this root is not admissible because the fourth moment used in the update for $q(\mu)$ is then undefined.

Proof: Consider the set of consistency conditions as the set of equations (72)-(75). Matching $\mathbb{V}_r(\boldsymbol{\mu})$ with $\mathbb{V}_q(\boldsymbol{\mu})$ leads to the equation

$$\frac{\tilde{\boldsymbol{\Psi}}}{\lambda_n(\tilde{d} - p - 1)} = \frac{\tilde{\nu}\tilde{\boldsymbol{\Sigma}}}{\tilde{\nu} - 2}.$$

Equating $\mathbb{E}_r(\boldsymbol{\Sigma})$ with $\mathbb{E}_q(\boldsymbol{\Sigma})$ and the above equation we get

$$\frac{\tilde{\boldsymbol{\Psi}}}{\tilde{d} - p - 1} = \frac{1}{\nu_n - p} \left[\boldsymbol{\Psi}_n + \frac{\tilde{\boldsymbol{\Psi}}}{\tilde{d} - p - 1} \right].$$

Solving for $\tilde{\boldsymbol{\Psi}}$ we obtain

$$\tilde{\boldsymbol{\Psi}} = \frac{\tilde{d} - p - 1}{\nu_n - p - 1} \boldsymbol{\Psi}_n.$$

Hence, we can write the expressions for $\mathbf{A}$ and $\text{dg}(\mathbf{B})$ as

$$\mathbf{A} = \boldsymbol{\Psi}_n + \frac{\tilde{\boldsymbol{\Psi}}}{(\tilde{d} - p - 1)} = \frac{\nu_n - p}{\tilde{d} - p - 1} \tilde{\boldsymbol{\Psi}}$$

and

$$\begin{aligned} \text{dg}(\mathbf{B}) &= \frac{\lambda_n^2 \tilde{\nu}^2 (\tilde{\nu} - 1)}{(\tilde{\nu} - 2)^2 (\tilde{\nu} - 4)} 2\text{dg}(\tilde{\boldsymbol{\Sigma}})^2 \\ &= \frac{1}{(\tilde{d} - p - 1)^2} \frac{(\tilde{\nu} - 1)}{(\tilde{\nu} - 4)} 2\text{dg}(\tilde{\boldsymbol{\Psi}})^2 \\ &= \frac{(\tilde{\nu} - 1)}{(\nu_n - p)^2 (\tilde{\nu} - 4)} 2\text{dg}(\mathbf{A})^2. \end{aligned}$$

Hence, the moment matching equation for $\tilde{d}$ can be written as

$$\begin{aligned} \tilde{d} &= \frac{2\mathbf{1}^T[\text{dg}(\mathbb{E}_r(\boldsymbol{\Sigma}))^2]}{\text{tr}[\mathbb{E}\mathbb{W}\mathbb{V}_r(\boldsymbol{\Sigma})]} + p + 3 \\ &= \frac{\frac{2}{(\nu_n - p)^2} \mathbf{1}^T[\text{dg}(\mathbf{A})^2]}{\frac{2\mathbf{1}^T[\text{dg}(\mathbf{A})^2] + (\nu_n - p)\text{tr}(\mathbf{B})}{(\nu_n - p)^2(\nu_n - p - 2)}} + p + 3 \\ &= (\nu_n - p - 2) \frac{2\mathbf{1}^T[\text{dg}(\mathbf{A})^2]}{2\mathbf{1}^T[\text{dg}(\mathbf{A})^2] + (\nu_n - p)\text{tr}(\mathbf{B})} + p + 3. \end{aligned}$$

Substituting the above expressions for $\mathbf{A}$ in terms of ν_n , $\tilde{d}$, p and $\tilde{\Psi}$, we have

$$\begin{aligned}
\tilde{d} &= \frac{(\nu_n - p - 2)}{1 + \frac{(\tilde{\nu} - 1)}{(\nu_n - p)(\tilde{\nu} - 4)}} + p + 3 \\
&= (\nu_n - p - 2) \frac{(\nu_n - p)(\tilde{\nu} - 4)}{(\nu_n - p)(\tilde{\nu} - 4) + (\tilde{\nu} - 1)} + p + 3 \\
&= (\nu_n - p - 2) \left[1 - \frac{(\tilde{\nu} - 1)}{(\nu_n - p)(\tilde{\nu} - 4) + (\tilde{\nu} - 1)} \right] + p + 3 \\
&= \nu_n + 1 - \frac{(\tilde{\nu} - 1)(\nu_n - p - 2)}{(\nu_n - p)(\tilde{\nu} - 4) + (\tilde{\nu} - 1)} \\
&= \nu_n + 1 - \frac{(\tilde{d} - p)(\nu_n - p - 2)}{(\nu_n - p)(\tilde{d} - p - 3) + (\tilde{d} - p)}
\end{aligned}$$

where the last line follows from using $\tilde{\nu} = \tilde{d} - p + 1$. Letting $x = \tilde{d} - p$ and $y = \nu_n - p$ leads to the equation

$$\begin{aligned}
x &= y + 1 - \frac{x(y - 2)}{y(x - 3) + x} \\
&= y + 1 - \frac{x(y + 1) - 3x}{x(y + 1) - 3y}.
\end{aligned}$$

Multiplying by $x(y + 1) - 3y$, expanding, simplifying, and grouping by powers of x , and dividing through by $y + 1$ we have

$$x^2 - (y + 3)x + 3y = 0$$

The solutions are:

$$x = \frac{y + 3 \pm (y - 3)}{2} = y \text{ or } 3.$$

The solution $x = y$ gives $\tilde{d} = \nu_n$, and hence the exact posterior parameters. The second algebraic root $x = 3$ corresponds to $\tilde{d} = p + 3$ and $\tilde{\nu} = 4$. This root lies on the boundary of the admissible parameter space, since the fourth moment of $t_p(\tilde{\mu}, \tilde{\Sigma}, \tilde{\nu})$ exists only for $\tilde{\nu} > 4$. Therefore, the exact-posterior fixed point is the unique admissible fixed point.

E.7 Example

Suppose we draw $n = 4$ samples from $\mathbf{x}_i \stackrel{\text{iid}}{\sim} N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix}$$

with summary statistics

$$\bar{\mathbf{x}} = \begin{bmatrix} -0.9724726 \\ 1.3202681 \end{bmatrix} \quad \text{and} \quad \mathbf{S} = \begin{bmatrix} 0.8144316 & 0.5688416 \\ 0.5688416 & 1.9682059 \end{bmatrix}.$$

Again, we have chosen the sample size to be small in order to see the biggest differences between MFVB and Exact methods.

The fitted values of the q -densities are summarised in Table 4. Figure 1 displays the posterior density estimates for MFVB against the exact posterior distributions. The posterior covariances for $\boldsymbol{\mu}$ are given by

$$\mathbb{E}_q^{VB}(\boldsymbol{\mu}) = \begin{bmatrix} 0.065 & 0.020 \\ 0.020 & 0.106 \end{bmatrix} \quad \text{and} \quad \mathbb{E}(\boldsymbol{\mu}|\mathbf{X}) = \begin{bmatrix} 0.114 & 0.035 \\ 0.035 & 0.186 \end{bmatrix},$$

and the element-wise variances for $\boldsymbol{\Sigma}$ are given by

$$\mathbb{E}_q^{VB}(\boldsymbol{\Sigma}) = \begin{bmatrix} 0.116 & 0.085 \\ 0.085 & 0.310 \end{bmatrix} \quad \text{and} \quad \mathbb{E}(\boldsymbol{\Sigma}|\mathbf{X}) = \begin{bmatrix} 0.208 & 0.148 \\ 0.148 & 0.557 \end{bmatrix},$$

showing the exact posterior variances for $\boldsymbol{\mu}$ are 1.75 times bigger than MFVB estimates, and that the MFVB estimates, and the exact posterior element-wise variances for $\boldsymbol{\Sigma}$ are 1.73-1.79 times bigger than the corresponding MFVB estimates.

Method	$\tilde{\nu}$	$\tilde{d}$	$\tilde{\Sigma}_{11}$	$\tilde{\Sigma}_{12}$	$\tilde{\Sigma}_{22}$	$\tilde{\Psi}_{11}$	$\tilde{\Psi}_{12}$	$\tilde{\Psi}_{22}$	Iterations
MFVB		8	0.065	0.0198	0.106	2.08	0.635	3.41	9
MP	6	7	0.114	0.0347	0.186	1.82	0.556	2.99	22
Exact	6	7	0.114	0.0347	0.186	1.82	0.556	2.99	

Table 4: Summary of fits for the MVN model using MFVB and MP methods for the described data. The values for $\tilde{\mu}$ are not shown because they are exact for all methods.

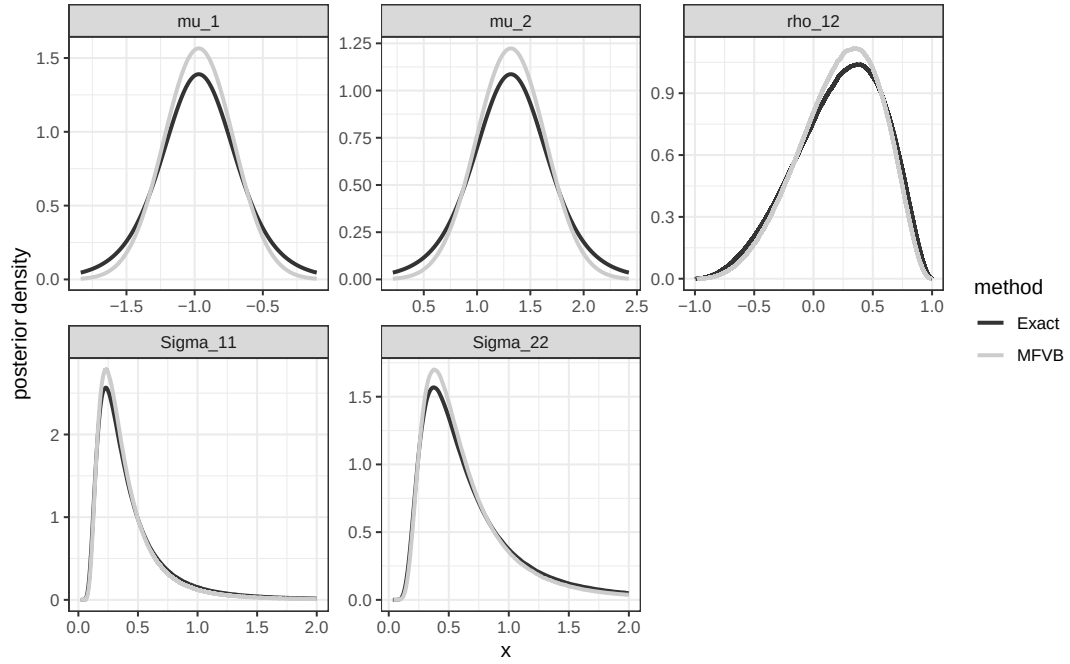


Figure 1: The fitted posteriors for the MVN model using described data. The posteriors densities corresponding to MP are not plotted because they are exact.

F Derivations for MVN model with missing values

We use the notation setup in Section 4.3 which we spell out more explicitly here. Let $\mathbf{X}$ be the $n \times p$ matrix whose i -th rows are equal to $\mathbf{x}_i^T$ and $\mathbf{X}_M$ denote the collection of missing values in the matrix $\mathbf{X}$. For $i \in \mathcal{M}$, write

$$\mathbf{z}_i = \mathbf{x}_{i,M_i}, \quad \mathbf{x}_i = \mathbf{c}_i + \mathbf{R}_i \mathbf{z}_i,$$

where O_i and M_i are index sets representing the elements of $\mathbf{x}_i$ that are observed and missing respectively, $\mathbf{c}_i$ is the vector obtained by replacing the missing entries of $\mathbf{x}_i$ by zero, and $\mathbf{R}_i$ is the matrix that inserts $\mathbf{z}_i$ into the missing positions. For example, if $p = 5$ and $M_i = \{2, 4, 5\}$ then

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ x_{i4} \\ x_{i5} \end{bmatrix}, \quad \mathbf{c}_i = \begin{bmatrix} x_{i1} \\ 0 \\ x_{i3} \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{R}_i = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{z}_i = \mathbf{x}_{i,M_i} = \begin{bmatrix} x_{i2} \\ x_{i4} \\ x_{i5} \end{bmatrix}.$$

F.1 Fully factorized mean-field variational Bayes

We first derive the MFVB updates corresponding to the factorization

$$q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X}_M) = q(\boldsymbol{\mu})q(\boldsymbol{\Sigma}) \prod_{i \in \mathcal{M}} q(\mathbf{x}_{i,M_i}),$$

where $\mathcal{M}$ denotes the set of observations with at least one missing component. We take

$$q(\boldsymbol{\mu}) = N_p(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \quad q(\boldsymbol{\Sigma}) = \text{IW}_p(\tilde{\boldsymbol{\Psi}}, \tilde{d}),$$

and, for $i \in \mathcal{M}$, $q(\mathbf{z}_i) = q(\mathbf{x}_{i,M_i}) = N_{|M_i|}(\tilde{\mathbf{m}}_i, \tilde{\mathbf{V}}_i)$.

Update for $q(\mathbf{x}_{i,M_i})$. For fixed $q(\boldsymbol{\mu})$ and $q(\boldsymbol{\Sigma})$, the terms in the log joint density involving $\mathbf{z}_i = \mathbf{x}_{i,M_i}$ are

$$-\frac{1}{2}(\mathbf{c}_i + \mathbf{R}_i \mathbf{z}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{c}_i + \mathbf{R}_i \mathbf{z}_i - \boldsymbol{\mu}).$$

Taking expectation with respect to $q(\boldsymbol{\mu})q(\boldsymbol{\Sigma})$, and retaining only terms depending on $\mathbf{z}_i$, gives

$$\log q^*(\mathbf{z}_i) = -\frac{1}{2} \mathbf{z}_i^T \mathbf{R}_i^T \tilde{\boldsymbol{\Omega}} \mathbf{R}_i \mathbf{z}_i + \mathbf{z}_i^T \mathbf{R}_i^T \tilde{\boldsymbol{\Omega}} (\tilde{\boldsymbol{\mu}} - \mathbf{c}_i) + \text{const.}$$

where $\tilde{\boldsymbol{\Omega}} = \mathbb{E}_q(\boldsymbol{\Sigma}^{-1}) = \tilde{d} \tilde{\boldsymbol{\Psi}}^{-1}$. Therefore,

$$q^*(\mathbf{z}_i) = N_{|M_i|}(\tilde{\mathbf{m}}_i, \tilde{\mathbf{V}}_i), \quad \text{where} \quad \tilde{\mathbf{V}}_i = (\mathbf{R}_i^T \tilde{\boldsymbol{\Omega}} \mathbf{R}_i)^{-1}, \quad \text{and} \quad \tilde{\mathbf{m}}_i = \tilde{\mathbf{V}}_i \mathbf{R}_i^T \tilde{\boldsymbol{\Omega}} (\tilde{\boldsymbol{\mu}} - \mathbf{c}_i).$$

Note that this can be equivalently written as follows. If $O_i \neq \emptyset$, this can be written in block form as $\tilde{\mathbf{V}}_i = \tilde{\boldsymbol{\Omega}}_{M_i M_i}^{-1}$, and $\tilde{\mathbf{m}}_i = \tilde{\boldsymbol{\mu}}_{M_i} - \tilde{\boldsymbol{\Omega}}_{M_i M_i}^{-1} \tilde{\boldsymbol{\Omega}}_{M_i O_i} (\mathbf{x}_{i,O_i} - \tilde{\boldsymbol{\mu}}_{O_i})$. If $O_i = \emptyset$, then the update reduces to $\tilde{\mathbf{m}}_i = \tilde{\boldsymbol{\mu}}$, and $\tilde{\mathbf{V}}_i = \tilde{\boldsymbol{\Omega}}^{-1}$.

It is useful to define the first two moments of the completed observation

$$\hat{\mathbf{x}}_i = \mathbb{E}_q(\mathbf{x}_i) = \mathbf{c}_i + \mathbf{R}_i \tilde{\mathbf{m}}_i, \quad \text{and} \quad \mathbf{B}_i = \mathbb{V}_q(\mathbf{x}_i) = \mathbf{R}_i \tilde{\mathbf{V}}_i \mathbf{R}_i^T.$$

For observations with no missing components, we set $\hat{\mathbf{x}}_i = \mathbf{x}_i$ and $\mathbf{B}_i = \mathbf{0}$.

Update for $q(\boldsymbol{\mu})$. For fixed $q(\boldsymbol{\Sigma})$ and $\{q(\mathbf{x}_{i,M_i}) : i \in \mathcal{M}\}$, the terms in the log joint density involving $\boldsymbol{\mu}$ are

$$-\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) - \frac{\lambda_0}{2} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}.$$

Taking expectation with respect to $q(\boldsymbol{\Sigma}) \prod_{i \in \mathcal{M}} q(\mathbf{x}_{i,M_i})$, and retaining only terms depending on $\boldsymbol{\mu}$, gives

$$\log q^*(\boldsymbol{\mu}) = -\frac{1}{2} \lambda_n \boldsymbol{\mu}^T \tilde{\boldsymbol{\Omega}} \boldsymbol{\mu} + \boldsymbol{\mu}^T \tilde{\boldsymbol{\Omega}} \sum_{i=1}^n \hat{\mathbf{x}}_i + \text{const.},$$

where $\lambda_n = \lambda_0 + n$. Thus, $q^*(\boldsymbol{\mu}) = N_p(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, with

$$\tilde{\boldsymbol{\mu}} = \frac{1}{\lambda_n} \sum_{i=1}^n \hat{\mathbf{x}}_i \equiv \hat{\boldsymbol{\mu}}_n, \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}} = (\lambda_n \tilde{\boldsymbol{\Omega}})^{-1}.$$

Update for $q(\mathbf{\Sigma})$. For fixed $q(\boldsymbol{\mu})$ and $\{q(\mathbf{x}_{i,M_i}) : i \in \mathcal{M}\}$, the terms in the log joint density involving $\mathbf{\Sigma}$ are

$$-\frac{\nu_0 + p + n + 2}{2} \log |\mathbf{\Sigma}| - \frac{1}{2} \text{tr}(\mathbf{\Psi}_0 \mathbf{\Sigma}^{-1}) - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) - \frac{\lambda_0}{2} \boldsymbol{\mu}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}.$$

Hence, $q^*(\mathbf{\Sigma}) = \text{IW}_p(\tilde{\mathbf{\Psi}}, \tilde{d})$, where

$$\tilde{d} = \nu_0 + n + 1 \quad \text{and} \quad \tilde{\mathbf{\Psi}} = \mathbf{\Psi}_0 + \sum_{i=1}^n \mathbb{E}_q \left[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \right] + \lambda_0 \mathbb{E}_q(\boldsymbol{\mu} \boldsymbol{\mu}^T).$$

Since $q(\boldsymbol{\mu})$ is independent of $q(\mathbf{x}_{i,M_i})$, we have

$$\mathbb{E}_q \left[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \right] = \mathbf{B}_i + \tilde{\mathbf{\Sigma}} + (\hat{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}})(\hat{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}})^T, \quad \text{and} \quad \mathbb{E}_q(\boldsymbol{\mu} \boldsymbol{\mu}^T) = \tilde{\mathbf{\Sigma}} + \tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\mu}}^T.$$

Therefore,

$$\tilde{\mathbf{\Psi}} = \mathbf{\Psi}_0 + \sum_{i=1}^n \left[\mathbf{B}_i + (\hat{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}})(\hat{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}})^T \right] + \lambda_0 \tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\mu}}^T + \lambda_n \tilde{\mathbf{\Sigma}}.$$

At the update for $q(\boldsymbol{\mu})$, we have $\tilde{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_n$, where $\hat{\boldsymbol{\mu}}_n = \lambda_n^{-1} \sum_{i=1}^n \hat{\mathbf{x}}_i$. Thus the scale update can equivalently be written as

$$\tilde{\mathbf{\Psi}} = \mathbf{\Psi}_0 + \sum_{i=1}^n \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^T + \sum_{i=1}^n \mathbf{B}_i - \lambda_n \hat{\boldsymbol{\mu}}_n \hat{\boldsymbol{\mu}}_n^T + \lambda_n \tilde{\mathbf{\Sigma}}.$$

This expression highlights the difference between the fully factorized MFVB update and the blocked update for $q(\boldsymbol{\mu}, \mathbf{\Sigma})$: the factorization $q(\boldsymbol{\mu})q(\mathbf{\Sigma})$ contributes the extra term $\lambda_n \tilde{\mathbf{\Sigma}}$ to the inverse-Wishart scale matrix and changes the degrees of freedom from $\nu_0 + n$ to $\nu_0 + n + 1$.

F.2 Blocked mean-field variational Bayes

We now derive the MFVB updates corresponding to the partially factorized approximation

$$q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X}_M) = q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i \in \mathcal{M}} q(\mathbf{x}_{i, M_i}).$$

We take

$$q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \text{NIW}_p(\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\boldsymbol{\Psi}}, \tilde{\nu}),$$

meaning that $q(\boldsymbol{\mu} | \boldsymbol{\Sigma}) = N_p(\tilde{\boldsymbol{\mu}}, \tilde{\lambda}^{-1} \boldsymbol{\Sigma})$, and $q(\boldsymbol{\Sigma}) = \text{IW}_p(\tilde{\boldsymbol{\Psi}}, \tilde{\nu})$, and, for $i \in \mathcal{M}$, $q(\mathbf{z}_i) = q(\mathbf{x}_{i, M_i}) = N_{|M_i|}(\tilde{\mathbf{m}}_i, \tilde{\mathbf{V}}_i)$. Define $\tilde{\boldsymbol{\Omega}} = \mathbb{E}_q(\boldsymbol{\Sigma}^{-1}) = \tilde{\nu} \tilde{\boldsymbol{\Psi}}^{-1}$.

Update for $q(\mathbf{x}_{i, M_i})$. For fixed $q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the terms in the log joint density involving $\mathbf{z}_i = \mathbf{x}_{i, M_i}$ are

$$-\frac{1}{2}(\mathbf{c}_i + \mathbf{R}_i \mathbf{z}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{c}_i + \mathbf{R}_i \mathbf{z}_i - \boldsymbol{\mu}).$$

Taking expectation with respect to $q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and retaining only terms depending on $\mathbf{z}_i$, gives

$$\log q^*(\mathbf{z}_i) = -\frac{1}{2} \mathbf{z}_i^T \mathbf{R}_i^T \tilde{\boldsymbol{\Omega}} \mathbf{R}_i \mathbf{z}_i + \mathbf{z}_i^T \mathbf{R}_i^T \tilde{\boldsymbol{\Omega}} (\tilde{\boldsymbol{\mu}} - \mathbf{c}_i) + \text{const}.$$

Therefore, $q^*(\mathbf{z}_i) = N_{|M_i|}(\tilde{\mathbf{m}}_i, \tilde{\mathbf{V}}_i)$, where

$$\tilde{\mathbf{V}}_i = (\mathbf{R}_i^T \tilde{\boldsymbol{\Omega}} \mathbf{R}_i)^{-1}, \quad \text{and} \quad \tilde{\mathbf{m}}_i = \tilde{\mathbf{V}}_i \mathbf{R}_i^T \tilde{\boldsymbol{\Omega}} (\tilde{\boldsymbol{\mu}} - \mathbf{c}_i).$$

Equivalently, if $O_i \neq \emptyset$, this update can be written in block form as

$$\tilde{\mathbf{V}}_i = \tilde{\boldsymbol{\Omega}}_{M_i M_i}^{-1}, \quad \text{and} \quad \tilde{\mathbf{m}}_i = \tilde{\boldsymbol{\mu}}_{M_i} - \tilde{\boldsymbol{\Omega}}_{M_i M_i}^{-1} \tilde{\boldsymbol{\Omega}}_{M_i O_i} (\mathbf{x}_{i, O_i} - \tilde{\boldsymbol{\mu}}_{O_i}).$$

If $O_i = \emptyset$, this reduces to $\tilde{\mathbf{m}}_i = \tilde{\boldsymbol{\mu}}$, and $\tilde{\mathbf{V}}_i = \tilde{\boldsymbol{\Omega}}^{-1}$. After updating $q(z_i)$, define

$$\hat{\mathbf{x}}_i = \mathbb{E}_q(x_i) = \mathbf{c}_i + \mathbf{R}_i \tilde{\mathbf{m}}_i, \quad \text{and} \quad \mathbf{B}_i = \mathbb{V}_q(x_i) = \mathbf{R}_i \tilde{\mathbf{V}}_i \mathbf{R}_i^T.$$

As previously, for observations with no missing components, we set $\hat{\mathbf{x}}_i = \mathbf{x}_i$, and $\mathbf{B}_i = \mathbf{0}$.

Update for $q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For fixed $\{q(\mathbf{x}_{i,M_i}) : i \in \mathcal{M}\}$, the terms in the log joint density involving $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are $\log p(\boldsymbol{\Sigma}) + \log p(\boldsymbol{\mu}|\boldsymbol{\Sigma}) + \sum_{i=1}^n \mathbb{E}_q[\log p(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})]$. Ignoring constants not depending on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, this is

$$-\frac{\nu_0 + n + p + 2}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \text{tr}(\boldsymbol{\Psi}_0 \boldsymbol{\Sigma}^{-1}) - \frac{\lambda_0}{2} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \frac{1}{2} \sum_{i=1}^n \mathbb{E}_q \left[(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right].$$

Since $\mathbb{E}_q(\mathbf{x}_i) = \hat{\mathbf{x}}_i$, and $\mathbb{V}_q(\mathbf{x}_i) = \mathbf{B}_i$, we have

$$\mathbb{E}_q \left[(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] = \text{tr} \left[\boldsymbol{\Sigma}^{-1} \left\{ \mathbf{B}_i + (\hat{\mathbf{x}}_i - \boldsymbol{\mu})(\hat{\mathbf{x}}_i - \boldsymbol{\mu})^T \right\} \right].$$

Therefore, the coordinate update for $q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is proportional to

$$|\boldsymbol{\Sigma}|^{-(\nu_0 + n + p + 2)/2} \exp \left[-\frac{1}{2} \text{tr} \left\{ \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\Psi}_0 + \sum_{i=1}^n \mathbf{B}_i + \sum_{i=1}^n (\hat{\mathbf{x}}_i - \boldsymbol{\mu})(\hat{\mathbf{x}}_i - \boldsymbol{\mu})^T + \lambda_0 \boldsymbol{\mu} \boldsymbol{\mu}^T \right) \right\} \right].$$

Let $\lambda_n = \lambda_0 + n$, $\bar{\mathbf{x}} = \sum_{i=1}^n \hat{\mathbf{x}}_i / n$, $\hat{\boldsymbol{\mu}}_n = n \bar{\mathbf{x}} / \lambda_n = \lambda_n^{-1} \sum_{i=1}^n \hat{\mathbf{x}}_i$. Completing the square in $\boldsymbol{\mu}$ gives

$$\sum_{i=1}^n (\hat{\mathbf{x}}_i - \boldsymbol{\mu})(\hat{\mathbf{x}}_i - \boldsymbol{\mu})^T + \lambda_0 \boldsymbol{\mu} \boldsymbol{\mu}^T = \lambda_n (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_n)(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_n)^T + \sum_{i=1}^n \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^T - \lambda_n \hat{\boldsymbol{\mu}}_n \hat{\boldsymbol{\mu}}_n^T.$$

Hence, $q^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \text{NIW}_p(\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\boldsymbol{\Psi}}, \tilde{\nu})$, where

$$\tilde{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_n, \quad \tilde{\lambda} = \lambda_n, \quad \tilde{\nu} = \nu_0 + n, \quad \text{and} \quad \tilde{\boldsymbol{\Psi}} = \boldsymbol{\Psi}_0 + \sum_{i=1}^n \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^T + \sum_{i=1}^n \mathbf{B}_i - \lambda_n \hat{\boldsymbol{\mu}}_n \hat{\boldsymbol{\mu}}_n^T.$$

Equivalently, $\tilde{\boldsymbol{\Psi}} = \boldsymbol{\Psi}_0 + \sum_{i=1}^n \left[\mathbf{B}_i + (\hat{\mathbf{x}}_i - \hat{\boldsymbol{\mu}}_n)(\hat{\mathbf{x}}_i - \hat{\boldsymbol{\mu}}_n)^T \right] + \lambda_0 \hat{\boldsymbol{\mu}}_n \hat{\boldsymbol{\mu}}_n^T$.

The blocked approximation differs from the fully factorized MFVB approximation in two ways. First, the variational distribution preserves the conditional dependence

$$q(\boldsymbol{\mu}|\boldsymbol{\Sigma}) = N_p(\tilde{\boldsymbol{\mu}}, \tilde{\lambda}^{-1} \boldsymbol{\Sigma}).$$

Second, the update for $q(\boldsymbol{\Sigma})$ has degrees of freedom $\nu_0 + n$, rather than $\nu_0 + n + 1$, because $\boldsymbol{\mu}$ is updated jointly with $\boldsymbol{\Sigma}$ rather than factorized from it.

F.3 Mean-field moment propagation

We now derive the MFMP updates for the factorization

$$q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X}_M) = q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i \in \mathcal{M}} q(\mathbf{z}_i),$$

where $\mathbf{z}_i = \mathbf{x}_{i, M_i}$ denotes the missing components of observation i . We use

$$q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \text{NIW}_p(\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\boldsymbol{\Psi}}, \tilde{\nu}), \quad \text{and} \quad q(\mathbf{z}_i) = N_{|M_i|}(\tilde{\mathbf{m}}_i, \tilde{\mathbf{V}}_i), \quad i \in \mathcal{M}.$$

The update for $q(\mathbf{z}_i)$ is obtained by propagating the conditional moments of $\mathbf{x}_{i, M_i}$ over the current approximation $q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The update for $q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is then obtained by propagating the conditional posterior moments of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ over the current approximation to the missing values.

Update for $q(\mathbf{z}_i)$. First suppose $O_i = \emptyset$. Then $\mathbf{z}_i = \mathbf{x}_i$, and conditional on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have $\mathbf{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Therefore, $\mathbb{E}_r(\mathbf{z}_i) = \mathbb{E}_q(\boldsymbol{\mu}) = \tilde{\boldsymbol{\mu}}$. Also, by the law of total variance, $\mathbb{V}_r(\mathbf{z}_i) = \mathbb{E}_q(\boldsymbol{\Sigma}) + \mathbb{V}_q(\boldsymbol{\mu})$. Under the normal-inverse-Wishart approximation, $\mathbb{V}_q(\boldsymbol{\mu}) = \tilde{\lambda}^{-1} \mathbb{E}_q(\boldsymbol{\Sigma})$, and hence

$$\mathbb{V}_r(\mathbf{z}_i) = (1 + \tilde{\lambda}^{-1}) \mathbb{E}_q(\boldsymbol{\Sigma}) = (1 + \tilde{\lambda}^{-1}) \frac{\tilde{\boldsymbol{\Psi}}}{\tilde{\nu} - p - 1}.$$

Thus, if $O_i = \emptyset$, the moment-matching update is

$$\tilde{\mathbf{m}}_i \leftarrow \tilde{\boldsymbol{\mu}}, \quad \text{and} \quad \tilde{\mathbf{V}}_i \leftarrow (1 + \tilde{\lambda}^{-1}) \frac{\tilde{\boldsymbol{\Psi}}}{\tilde{\nu} - p - 1}.$$

Now suppose $O_i \neq \emptyset$. Conditional on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the missing components have Gaussian conditional distribution

$$\mathbf{x}_{i, M_i} | \mathbf{x}_{i, O_i}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_{|M_i|} \left(\boldsymbol{\mu}_{M_i} + \boldsymbol{\Sigma}_{M_i O_i} \boldsymbol{\Sigma}_{O_i O_i}^{-1} (\mathbf{x}_{i, O_i} - \boldsymbol{\mu}_{O_i}), \boldsymbol{\Sigma}_{M_i | O_i} \right),$$

where $\boldsymbol{\Sigma}_{M_i | O_i} = \boldsymbol{\Sigma}_{M_i M_i} - \boldsymbol{\Sigma}_{M_i O_i} \boldsymbol{\Sigma}_{O_i O_i}^{-1} \boldsymbol{\Sigma}_{O_i M_i}$. Define $\mathbf{L}_i = \mathbb{E}_q(\boldsymbol{\Sigma}_{M_i O_i} \boldsymbol{\Sigma}_{O_i O_i}^{-1}) = \tilde{\boldsymbol{\Psi}}_{M_i O_i} \tilde{\boldsymbol{\Psi}}_{O_i O_i}^{-1}$ (see Section A.4), and $\tilde{\boldsymbol{\Psi}}_{M_i | O_i} = \tilde{\boldsymbol{\Psi}}_{M_i M_i} - \tilde{\boldsymbol{\Psi}}_{M_i O_i} \tilde{\boldsymbol{\Psi}}_{O_i O_i}^{-1} \tilde{\boldsymbol{\Psi}}_{O_i M_i}$. Then $\mathbb{E}_r(\mathbf{z}_i) = \tilde{\boldsymbol{\mu}}_{M_i} + \mathbf{L}_i (\mathbf{x}_{i, O_i} - \tilde{\boldsymbol{\mu}}_{O_i})$.

The propagated variance is obtained using the law of total variance. We have

$$\mathbb{V}_r(\mathbf{z}_i) = \mathbb{E}_q(\boldsymbol{\Sigma}_{M_i|O_i}) + \mathbb{V}_q \left[\boldsymbol{\mu}_{M_i} + \boldsymbol{\Sigma}_{M_i O_i} \boldsymbol{\Sigma}_{O_i O_i}^{-1} (\mathbf{x}_{i,O_i} - \boldsymbol{\mu}_{O_i}) \right].$$

Using the block inverse-Wishart identities in Section A.4, this gives

$$\mathbb{V}_r(\mathbf{z}_i) = \left[1 + \tilde{\lambda}^{-1} + (\mathbf{x}_{i,O_i} - \tilde{\boldsymbol{\mu}}_{O_i})^T \tilde{\boldsymbol{\Psi}}_{O_i O_i}^{-1} (\mathbf{x}_{i,O_i} - \tilde{\boldsymbol{\mu}}_{O_i}) \right] \frac{\tilde{\boldsymbol{\Psi}}_{M_i|O_i}}{\tilde{\nu} - |M_i| - 1}.$$

Hence, for $O_i \neq \emptyset$,

$$\tilde{\mathbf{m}}_i \leftarrow \tilde{\boldsymbol{\mu}}_{M_i} + \mathbf{L}_i (\mathbf{x}_{i,O_i} - \tilde{\boldsymbol{\mu}}_{O_i}),$$

and

$$\tilde{\mathbf{V}}_i \leftarrow \left[1 + \tilde{\lambda}^{-1} + (\mathbf{x}_{i,O_i} - \tilde{\boldsymbol{\mu}}_{O_i})^T \tilde{\boldsymbol{\Psi}}_{O_i O_i}^{-1} (\mathbf{x}_{i,O_i} - \tilde{\boldsymbol{\mu}}_{O_i}) \right] \frac{\tilde{\boldsymbol{\Psi}}_{M_i|O_i}}{\tilde{\nu} - |M_i| - 1}.$$

After updating $q(\mathbf{z}_i)$, define $\hat{\mathbf{x}}_i = \mathbb{E}_q(\mathbf{x}_i) = \mathbf{c}_i + \mathbf{R}_i \tilde{\mathbf{m}}_i$ and $\mathbf{B}_i = \mathbb{V}_q(\mathbf{x}_i) = \mathbf{R}_i \tilde{\mathbf{V}}_i \mathbf{R}_i^T$. Again, for observations with no missing components, set $\hat{\mathbf{x}}_i = \mathbf{x}_i$, and $\mathbf{B}_i = \mathbf{0}$.

Update for $q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For fixed $\{q(\mathbf{z}_i) : i \in \mathcal{M}\}$, the conditional posterior of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given the completed data matrix is normal-inverse-Wishart. Let $\lambda_n = \lambda_0 + n$, $\nu_n = \nu_0 + n$, and define $\hat{\boldsymbol{\mu}}_n = \lambda_n^{-1} \sum_{i=1}^n \hat{\mathbf{x}}_i$, and $\mathbf{A} = \sum_{i=1}^n \mathbf{B}_i$. The propagated mean of $\boldsymbol{\mu}$ is $\mathbb{E}_r(\boldsymbol{\mu}) = \hat{\boldsymbol{\mu}}_n$. The propagated variance of $\boldsymbol{\mu}$ is

$$\mathbb{V}_r(\boldsymbol{\mu}) = \frac{\mathbb{E}_r(\boldsymbol{\Sigma})}{\lambda_n} + \frac{\mathbf{A}}{\lambda_n^2}.$$

The second term is the contribution from uncertainty in the missing values.

The following moment calculations assume that the required inverse-Wishart moments exist; in particular, we require $\tilde{\nu} > p + 1$ for $\mathbb{E}_q(\boldsymbol{\Sigma})$ and $\nu_n > p + 3$ for the propagated element-wise variances of $\boldsymbol{\Sigma}$. The propagated mean of $\boldsymbol{\Sigma}$ is

$$\mathbb{E}_r(\boldsymbol{\Sigma}) = \frac{\mathbf{C}}{\nu_n - p - 1}$$

where $\mathbf{C} = \boldsymbol{\Psi}_0 + \sum_{i=1}^n \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^T + (1 - \lambda_n^{-1}) \mathbf{A} - \lambda_n \hat{\boldsymbol{\mu}}_n \hat{\boldsymbol{\mu}}_n^T$. To update the inverse-Wishart degrees

of freedom, we also need the diagonal element-wise variances of $\mathbf{\Sigma}$. Let

$$d_j = \mathbb{V}_q(\{\mathbf{\Psi}_n(X)\}_{jj}), \quad j = 1, \dots, p,$$

where

$$\mathbf{\Psi}_n(\mathbf{X}) = \mathbf{\Psi}_0 + \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T - \lambda_n \boldsymbol{\mu}_n(\mathbf{X}) \boldsymbol{\mu}_n(\mathbf{X})^T, \quad \text{and} \quad \boldsymbol{\mu}_n(\mathbf{X}) = \lambda_n^{-1} \sum_{i=1}^n \mathbf{x}_i.$$

Then

$$\mathbb{V}_r(\Sigma_{jj}) = \frac{2C_{jj}^2 + (\nu_n - p - 1)d_j}{(\nu_n - p - 1)^2(\nu_n - p - 3)}.$$

Equivalently, with $\mathbf{d} = (d_1, \dots, d_p)$,

$$\text{dg}[\mathbb{E}\mathbb{W}\mathbb{V}_r(\mathbf{\Sigma})] = \frac{2 \text{dg}(\mathbf{C})^{\odot 2} + (\nu_n - p - 1)\mathbf{d}}{(\nu_n - p - 1)^2(\nu_n - p - 3)}.$$

It remains to compute each of the d_j . Under the current approximation to the missing data, $\mathbf{x}_i$ has mean $\hat{\mathbf{x}}_i$ and covariance $\mathbf{B}_i$, independently across i . For a fixed coordinate j , define $\mathbf{X}_j = (x_{1j}, \dots, x_{nj})$, $\mathbf{a}_j = (\hat{x}_{1j}, \dots, \hat{x}_{nj})$, and $\mathbf{b}_j = (\{\mathbf{B}_1\}_{jj}, \dots, \{\mathbf{B}_n\}_{jj})$. Then $\mathbf{X}_j \sim N_n[\mathbf{a}_j, \text{diag}(\mathbf{b}_j)]$. Let $\mathbf{Q} = \mathbf{I}_n - \lambda_n^{-1} \mathbf{1}_n \mathbf{1}_n^T$. Ignoring the fixed prior term $\{\mathbf{\Psi}_0\}_{jj}$, the j -th diagonal element of $\mathbf{\Psi}_n(\mathbf{X})$ is $\mathbf{X}_j^T \mathbf{Q} \mathbf{X}_j$. Therefore, $d_j = \mathbb{V}_q(\mathbf{X}_j^T \mathbf{Q} \mathbf{X}_j)$. Using the Gaussian quadratic-form variance identity (see Section A.1),

$$d_j = 2 \text{tr}[\mathbf{Q} \text{diag}(\mathbf{b}_j) \mathbf{Q} \text{diag}(\mathbf{b}_j)] + 4 \mathbf{a}_j^T \mathbf{Q} \text{diag}(\mathbf{b}_j) \mathbf{Q} \mathbf{a}_j.$$

Equivalently,

$$d_j = 2 \left(1 - \frac{2}{\lambda_n}\right) \mathbf{1}_n^T \mathbf{b}_j^{\odot 2} + \frac{2}{\lambda_n^2} (\mathbf{1}_n^T \mathbf{b}_j)^2 + 4 \mathbf{b}_j^T \left(\mathbf{a}_j - \frac{\mathbf{1}_n^T \mathbf{a}_j}{\lambda_n} \mathbf{1}_n \right)^{\odot 2}.$$

Finally, we match the corresponding moments of the normal-inverse-Wishart approximation using

$$\mathbb{E}_q(\boldsymbol{\mu}) = \tilde{\boldsymbol{\mu}}, \quad \text{and} \quad \mathbb{V}_q(\boldsymbol{\mu}) = \frac{\mathbb{E}_q(\mathbf{\Sigma})}{\tilde{\lambda}},$$

Matching $\mathbb{E}_r(\boldsymbol{\mu}) = \mathbb{E}_q(\boldsymbol{\mu})$, $\text{tr}[\mathbb{V}_r(\boldsymbol{\mu})] = \text{tr}[\mathbb{V}_q(\boldsymbol{\mu})]$, $\mathbb{E}_r(\boldsymbol{\Sigma}) = \mathbb{E}_q(\boldsymbol{\Sigma})$, and $\mathbf{1}_p^T \text{dg}[\mathbb{E}\mathbb{W}\mathbb{V}_r(\boldsymbol{\Sigma})] = \mathbf{1}_p^T \text{dg}[\mathbb{E}\mathbb{W}\mathbb{V}_q(\boldsymbol{\Sigma})]$ and solving for $\tilde{\boldsymbol{\mu}}$, $\tilde{\lambda}$, $\tilde{\boldsymbol{\Psi}}$ and $\tilde{\nu}$ we obtain the updates:

$$\tilde{\boldsymbol{\mu}} \leftarrow \mathbb{E}_r(\boldsymbol{\mu}); \quad \tilde{\lambda} \leftarrow \frac{\text{tr}[\mathbb{E}_r(\boldsymbol{\Sigma})]}{\text{tr}[\mathbb{V}_r(\boldsymbol{\mu})]}; \quad \tilde{\nu} \leftarrow \frac{2 \mathbf{1}_p^T [\text{dg}(\mathbb{E}_r(\boldsymbol{\Sigma}))^{\odot 2}]}{\mathbf{1}_p^T \text{dg}[\mathbb{E}\mathbb{W}\mathbb{V}_r(\boldsymbol{\Sigma})]} + p + 3; \quad \tilde{\boldsymbol{\Psi}} \leftarrow (\tilde{\nu} - p - 1) \mathbb{E}_r(\boldsymbol{\Sigma}).$$

When there is no missing data, $\mathbf{A} = \mathbf{0}$ and $\mathbf{d} = \mathbf{0}$, and the update reduces to the complete-data normal-inverse-Wishart update. When missing data are present, $\mathbf{A}$ and $\mathbf{d}$ propagate uncertainty from the missing components into the moment-matching updates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Algorithm 6 summarizes the MFMP approximation.

Algorithm 6 Blocked MFMP algorithm for the MVN model with missing data

Require: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{\Psi}_0 \in \mathbb{S}_+^p$, $\lambda_0 > 0$, $\nu_0 + n > p + 3$.

- 1: Initialise: $\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\boldsymbol{\Psi}}$ and $\tilde{\nu}$ and set $\lambda_n \leftarrow \lambda_0 + n$; $\nu_n \leftarrow \nu_0 + n$;
- 2: For each i determine $\mathbf{c}_i, \mathbf{R}_i, O_i$ and M_i from $\mathbf{X}$ and set $\hat{\mathbf{x}}_i \leftarrow \mathbf{x}_i$, and $\mathbf{B}_i \leftarrow \mathbf{0}$ for all i s.t. $M_i = \emptyset$.
- 3: **repeat**
- 4: **for** i such that $M_i \neq \emptyset$ **do**
- 5: **if** $O_i = \emptyset$ **then**
- 6: Approximate posterior moments of $\mathbf{z}_i$ when $O_i = \emptyset$:

$$\mathbb{E}_r(\mathbf{z}_i) \leftarrow \tilde{\boldsymbol{\mu}}; \quad \mathbb{V}_r(\mathbf{z}_i) \leftarrow \left(1 + \frac{1}{\tilde{\lambda}}\right) \frac{\tilde{\boldsymbol{\Psi}}}{\tilde{\nu} - p - 1};$$

- 7: **else**
- 8: Approximate posterior moments of $\mathbf{z}_i$ when $O_i \neq \emptyset$:

$$\begin{aligned} \mathbf{L}_i &\leftarrow \tilde{\boldsymbol{\Psi}}_{M_i O_i} \tilde{\boldsymbol{\Psi}}_{O_i O_i}^{-1}; & \tilde{\boldsymbol{\Psi}}_{M_i | O_i} &\leftarrow \tilde{\boldsymbol{\Psi}}_{M_i M_i} - \tilde{\boldsymbol{\Psi}}_{M_i O_i} \tilde{\boldsymbol{\Psi}}_{O_i O_i}^{-1} \tilde{\boldsymbol{\Psi}}_{O_i M_i}; \\ \mathbb{E}_r(\mathbf{z}_i) &\leftarrow \tilde{\boldsymbol{\mu}}_{M_i} + \mathbf{L}_i(\mathbf{x}_{i, O_i} - \tilde{\boldsymbol{\mu}}_{O_i}); \\ \mathbb{V}_r(\mathbf{z}_i) &\leftarrow \left[1 + \tilde{\lambda}^{-1} + (\mathbf{x}_{i, O_i} - \tilde{\boldsymbol{\mu}}_{O_i})^T \tilde{\boldsymbol{\Psi}}_{O_i O_i}^{-1} (\mathbf{x}_{i, O_i} - \tilde{\boldsymbol{\mu}}_{O_i})\right] \frac{\tilde{\boldsymbol{\Psi}}_{M_i | O_i}}{\tilde{\nu} - |M_i| - 1}; \end{aligned}$$

- 9: **end if**
- 10: Perform MP update for $q(\mathbf{z}_i) = N(\tilde{\mathbf{m}}_i, \tilde{\mathbf{V}}_i)$ via moment matching

$$\tilde{\mathbf{m}}_i \leftarrow \mathbb{E}_r(\mathbf{z}_i); \quad \tilde{\mathbf{V}}_i \leftarrow \mathbb{V}_r(\mathbf{z}_i);$$

- 11: and update $\mathbb{E}_q(\mathbf{x}_i)$ and $\mathbb{V}_q(\mathbf{x}_i)$: $\hat{\mathbf{x}}_i \leftarrow \mathbf{c}_i + \mathbf{R}_i \tilde{\mathbf{m}}_i$; $\mathbf{B}_i \leftarrow \mathbf{R}_i \tilde{\mathbf{V}}_i \mathbf{R}_i^T$;
- 12: **end for**
- 13: Calculate intermediate quantities:

$$\hat{\boldsymbol{\mu}}_n \leftarrow \sum_{i=1}^n \hat{\mathbf{x}}_i / \lambda_n; \quad \mathbf{A} \leftarrow \sum_{i=1}^n \mathbf{B}_i; \quad \mathbf{C} \leftarrow \mathbf{\Psi}_0 + \sum_{i=1}^n \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^T + (1 - \lambda_n^{-1}) \mathbf{A} - \lambda_n \hat{\boldsymbol{\mu}}_n \hat{\boldsymbol{\mu}}_n^T$$

and for $j = 1, \dots, p$ set $\mathbf{a}_j = (\hat{x}_{1j}, \dots, \hat{x}_{nj})$ and $\mathbf{b}_j = (\{\mathbf{B}_1\}_{jj}, \dots, \{\mathbf{B}_n\}_{jj})$ and

$$d_j \leftarrow 2 \left(1 - \frac{2}{\lambda_n}\right) \mathbf{1}^T \mathbf{b}_j^{\odot 2} + \frac{2}{\lambda_n^2} (\mathbf{1}^T \mathbf{b}_j)^2 + 4 \mathbf{b}_j^T \left(\mathbf{a}_j - \frac{\mathbf{1}^T \mathbf{a}_j}{\lambda_n} \mathbf{1}_n\right)^{\odot 2};$$

- 14: Calculate local likelihood based target moments:

$$\begin{aligned} \mathbb{E}_r(\boldsymbol{\mu}) &\leftarrow \hat{\boldsymbol{\mu}}_n; \quad \mathbb{E}_r(\boldsymbol{\Sigma}) \leftarrow \frac{\mathbf{C}}{\nu_n - p - 1}; \quad \mathbb{V}_r(\boldsymbol{\mu}) \leftarrow \frac{\mathbb{E}_r(\boldsymbol{\Sigma})}{\lambda_n} + \frac{\mathbf{A}}{\lambda_n^2}; \\ \text{dg}[\mathbb{E} \mathbb{W} \mathbb{V}_r(\boldsymbol{\Sigma})] &\leftarrow \frac{2 \text{dg}(\mathbf{C})^{\odot 2} + (\nu_n - p - 1) \mathbf{d}}{(\nu_n - p - 1)^2 (\nu_n - p - 3)}; \end{aligned}$$

- 15: Set $\mathbb{E}_r(\boldsymbol{\mu}) = \mathbb{E}_q(\boldsymbol{\mu})$, $\text{tr}[\mathbb{V}_r(\boldsymbol{\mu})] = \text{tr}[\mathbb{V}_q(\boldsymbol{\mu})]$, $\mathbb{E}_r(\boldsymbol{\Sigma}) = \mathbb{E}_q(\boldsymbol{\Sigma})$, and $\mathbf{1}_p^T \text{dg}[\mathbb{E} \mathbb{W} \mathbb{V}_r(\boldsymbol{\Sigma})] = \mathbf{1}_p^T \text{dg}[\mathbb{E} \mathbb{W} \mathbb{V}_q(\boldsymbol{\Sigma})]$ to determine $\tilde{\boldsymbol{\mu}}, \tilde{\lambda}, \tilde{\boldsymbol{\Psi}}$ and $\tilde{\nu}$ via:

$$\tilde{\boldsymbol{\mu}} \leftarrow \mathbb{E}_r(\boldsymbol{\mu}); \quad \tilde{\lambda} \leftarrow \frac{\text{tr}[\mathbb{E}_r(\boldsymbol{\Sigma})]}{\text{tr}[\mathbb{V}_r(\boldsymbol{\mu})]}; \quad \tilde{\nu} \leftarrow \frac{2 \mathbf{1}_p^T [\text{dg}(\mathbb{E}_r(\boldsymbol{\Sigma}))^{\odot 2}]}{\mathbf{1}_p^T \text{dg}[\mathbb{E} \mathbb{W} \mathbb{V}_r(\boldsymbol{\Sigma})]} + p + 3; \quad \tilde{\boldsymbol{\Psi}} \leftarrow (\tilde{\nu} - p - 1) \mathbb{E}_r(\boldsymbol{\Sigma});$$

- 16: **until** convergence criteria are met
-

G Details for probit regression example

G.1 Derivations for MFVB

The derivation, except for very minor changes can be found in Ormerod and Wand (2010) (setting $\mathbf{y} = \mathbf{1}_n$, and using $\mathbf{Z} = \text{diag}(2\mathbf{y} - \mathbf{1}_n)\mathbf{X}$ in place of $\mathbf{X}$). We refer the interested reader there. The MFVB approximation corresponding to $q(\boldsymbol{\beta}, \mathbf{a}) = q(\boldsymbol{\beta})q(\mathbf{a})$ is

$$q(\boldsymbol{\beta}) = N(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \quad \text{and} \quad q(a_i) = \text{TN}_+(\tilde{m}_i, 1), \quad i = 1, \dots, n,$$

where the update for $q(\boldsymbol{\beta})$ is given by $\tilde{\boldsymbol{\mu}} \leftarrow \mathbf{S}\mathbf{Z}^T\mathbb{E}_q(\mathbf{a})$, and $\tilde{\boldsymbol{\Sigma}} \leftarrow \mathbf{S}$. Algorithm 7 summarizes the MFVB algorithm for the probit regression model. Algorithm 7 is very fast (per iteration) due to the fact that the matrix $\mathbf{S}$ needs to be only calculated once outside the main loop.

Algorithm 7 MFVB for probit regression

Require: $\mathbf{y} \in \{0, 1\}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, prior precision matrix $\mathbf{D}$, and initial value for $\tilde{\boldsymbol{\mu}}$.

- 1: Set $\mathbf{Z} = \text{diag}(2\mathbf{y} - \mathbf{1}_n)\mathbf{X}$, and $\mathbf{S} = (\mathbf{Z}^T\mathbf{Z} + \mathbf{D})^{-1}$
 - 2: **repeat**
 - 3: Update $q(\mathbf{a})$ via $\tilde{\mathbf{m}} \leftarrow \mathbf{Z}\tilde{\boldsymbol{\mu}}$
 - 4: Update $q(\boldsymbol{\beta})$ via $\tilde{\boldsymbol{\mu}} \leftarrow \mathbf{S}\mathbf{Z}^T(\tilde{\mathbf{m}} + \zeta_1(\tilde{\mathbf{m}}))$; $\tilde{\boldsymbol{\Sigma}} \leftarrow \mathbf{S}$;
 - 5: **until** convergence criteria are met
-

G.2 Additional details for MP

Derivations for the MP approximation for probit regression outline in Section 4.4 are relatively complete. A few details are included here for completeness.

Since $a_i|\beta, y_i$ is a unit-variance normal distribution truncated below at zero,

$$\mathbb{E}_q(a_i) = \tilde{m}_i + \zeta_1(\tilde{m}_i), \quad \text{where} \quad \zeta_k(t) = \frac{d^k}{dt^k} \log \Phi(t).$$

Algorithm 8 uses the function $\xi_k(m, v)$ which is defined as follows. For $k = 1, 2$,

$$\xi_k(m, v) = \mathbb{E}[\zeta_k(T)], \quad \text{where} \quad T \sim N(m, v),$$

with the convention that for vectors $\mathbf{m}$ and $\mathbf{v}$, $\xi_k(\mathbf{m}, \mathbf{v})$ is evaluated component-wise. The full algorithm is listed as Algorithm 8.

Algorithm 8 MP for probit regression

Require: $\mathbf{y} \in \{0, 1\}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, prior precision matrix $\mathbf{D}$, and initial value for $\tilde{\boldsymbol{\mu}}$.

- 1: Set $\mathbf{Z} = \text{diag}(2\mathbf{y} - \mathbf{1}_n)\mathbf{X}$, $\mathbf{S} = (\mathbf{Z}^T\mathbf{Z} + \mathbf{D})^{-1}$ and initialize $\tilde{\boldsymbol{\Sigma}} = \mathbf{S}$.
- 2: **repeat**
- 3: Calculate $\mathbf{Z}\tilde{\boldsymbol{\mu}}$ and $\text{dg}(\mathbf{Z}\tilde{\boldsymbol{\Sigma}}\mathbf{Z}^T)$
- 4: Approximate $\xi_1(\mathbf{Z}\tilde{\boldsymbol{\mu}}, \text{dg}(\mathbf{Z}\tilde{\boldsymbol{\Sigma}}\mathbf{Z}^T))$ and $\xi_2(\mathbf{Z}\tilde{\boldsymbol{\mu}}, \text{dg}(\mathbf{Z}\tilde{\boldsymbol{\Sigma}}\mathbf{Z}^T))$ via quadrature.
- 5: Update $q(\mathbf{a})$ via $\tilde{\mathbf{m}} \leftarrow \mathbf{Z}\tilde{\boldsymbol{\mu}} + \xi_1(\mathbf{Z}\tilde{\boldsymbol{\mu}}, \text{dg}(\mathbf{Z}\tilde{\boldsymbol{\Sigma}}\mathbf{Z}^T))$.
- 6: Update $q(\beta)$ via

$$\begin{aligned} \tilde{\boldsymbol{\mu}} &\leftarrow \mathbf{S}\mathbf{Z}^T\tilde{\mathbf{m}} \\ \tilde{\boldsymbol{\Sigma}} &\leftarrow \mathbf{S} + \mathbf{S}\mathbf{Z}^T [\mathbf{I}_n + \text{diag}\{\xi_2(\mathbf{Z}\tilde{\boldsymbol{\mu}}, \text{dg}(\mathbf{Z}\tilde{\boldsymbol{\Sigma}}\mathbf{Z}^T))\}] \mathbf{Z}\mathbf{S} \\ &\quad + \mathbf{S}\mathbf{Z}^T [\mathbf{I}_n + \text{diag}\{\zeta_2(\mathbf{Z}\tilde{\boldsymbol{\mu}})\}] \mathbf{Z}\tilde{\boldsymbol{\Sigma}}\mathbf{Z}^T [\mathbf{I}_n + \text{diag}\{\zeta_2(\mathbf{Z}\tilde{\boldsymbol{\mu}})\}] \mathbf{Z}\mathbf{S} \end{aligned}$$

- 7: **until** convergence criteria are met
-

G.3 Proof of Theorem 5

Regularity conditions. Assume that p is fixed and that the following conditions hold.

- (A1) The prior precision matrix $\mathbf{D}$ is fixed and positive definite.
- (A2) The posterior mode $\hat{\boldsymbol{\beta}}$ exists, is unique, and lies in the interior of $\mathbb{R}^p$. The MP fixed point $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ exists.
- (A3) The design matrix satisfies $n^{-1}\mathbf{X}^T\mathbf{X} = n^{-1}\mathbf{Z}^T\mathbf{Z} \rightarrow \boldsymbol{\Sigma}_x$, where $\boldsymbol{\Sigma}_x \in \mathcal{S}_p^+$. In particular, the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are of order n .
- (A4) At the MP fixed point, $\tilde{\boldsymbol{\Sigma}} = O_p^m(n^{-1})$, where $O_p^m(r_n)$ means that every element of the matrix is $O_p(r_n)$.
- (A5) Let $\mathbf{s}^2 = \text{dg}(\mathbf{Z}\tilde{\boldsymbol{\Sigma}}\mathbf{Z}^T)$. The Taylor series remainders in the Gaussian-smoothed probit derivatives satisfy $\mathbf{Z}^T\boldsymbol{\Delta}_1 = O_p^m(1)$, where $\xi_k(\mathbf{Z}\tilde{\boldsymbol{\mu}}, \mathbf{s}^2) = \zeta_k(\mathbf{Z}\tilde{\boldsymbol{\mu}}) + \boldsymbol{\Delta}_k$, $k = 1, 2$, and $\mathbf{Z}^T \text{diag}(\boldsymbol{\Delta}_2)\mathbf{Z} = O_p^m(1)$. In addition, if $\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\beta}} = O_p^m(n^{-1})$, assume $\mathbf{Z}^T[\text{diag}\{\zeta_2(\mathbf{Z}\tilde{\boldsymbol{\mu}})\} - \text{diag}\{\zeta_2(\mathbf{Z}\hat{\boldsymbol{\beta}})\}]\mathbf{Z} = O_p^m(1)$.
- (A6) For every $\bar{\boldsymbol{\beta}}$ on the line segment between $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\mu}}$, the matrix

$$\mathbf{D} - \mathbf{Z}^T \text{diag}[\zeta_2(\mathbf{Z}\bar{\boldsymbol{\beta}})]\mathbf{Z}$$

has eigenvalues of order n , bounded away from zero after scaling by n^{-1} .

Conditions (A3)–(A6) are the main order conditions used in the proof. Condition (A3) controls the scale of $\mathbf{Z}^T\mathbf{Z}$, while (A6) ensures that the posterior curvature matrix and the corresponding MP curvature matrix are invertible with inverse of order $O_p(n^{-1})$. Condition (A5) controls the Taylor remainders arising from replacing the Gaussian-smoothed quantities $\xi_j(\mathbf{Z}\tilde{\boldsymbol{\mu}}, \mathbf{s}^2)$ by their zero-variance limits $\zeta_j(\mathbf{Z}\tilde{\boldsymbol{\mu}})$. A simple sufficient condition for (A5) is a bounded-design assumption together with $\tilde{\boldsymbol{\Sigma}}^* = O_p^m(n^{-1})$. For random designs, suitable moment and tail conditions on the covariates also imply these bounds.

Before we begin the proof, the following two lemmas are needed.

Lemma 3 [Inverse perturbation bound]. *Let $\mathbf{A}_n$ and $\mathbf{E}_n$ be square matrices of fixed dimension. Suppose that $\mathbf{A}_n$ and $\mathbf{A}_n + \mathbf{E}_n$ are nonsingular, and that for any fixed matrix norm,*

$$\|\mathbf{A}_n^{-1}\| = O_p(a_n), \quad \|(\mathbf{A}_n + \mathbf{E}_n)^{-1}\| = O_p(a_n), \quad \|\mathbf{E}_n\| = O_p(b_n).$$

Then $(\mathbf{A}_n + \mathbf{E}_n)^{-1} - \mathbf{A}_n^{-1} = O_p(a_n^2 b_n)$. In particular, if $\mathbf{A}_n$ has eigenvalues of order n , so that $\|\mathbf{A}_n^{-1}\| = O_p(n^{-1})$, and if $\|\mathbf{E}_n\| = O_p(1)$, then

$$(\mathbf{A}_n + \mathbf{E}_n)^{-1} = \mathbf{A}_n^{-1} + O_p(n^{-2}).$$

Proof: First we use the exact identity $(\mathbf{A}_n + \mathbf{E}_n)^{-1} - \mathbf{A}_n^{-1} = -(\mathbf{A}_n + \mathbf{E}_n)^{-1} \mathbf{E}_n \mathbf{A}_n^{-1}$. Taking norms gives

$$\|(\mathbf{A}_n + \mathbf{E}_n)^{-1} - \mathbf{A}_n^{-1}\| \leq \|(\mathbf{A}_n + \mathbf{E}_n)^{-1}\| \|\mathbf{E}_n\| \|\mathbf{A}_n^{-1}\|.$$

By the assumed stochastic orders, the right-hand side is $O_p(a_n) O_p(b_n) O_p(a_n) = O_p(a_n^2 b_n)$. The special case follows by taking $a_n = n^{-1}$ and $b_n = 1$.

Lemma 4 [Inverse bound for the covariance fixed-point operator]. *Let $\mathbf{A}_n$, $\mathbf{B}_n$, and $\mathbf{C}_n$ be symmetric $p \times p$ matrices, where p is fixed. Suppose that $\mathbf{A}_n = \mathbf{B}_n + \mathbf{C}_n$, that $\mathbf{B}_n$ is positive semidefinite, and that $\mathbf{A}_n$ and $\mathbf{C}_n$ are positive definite with eigenvalues of order n . Define the linear operator*

$$\mathcal{L}_n(\mathbf{E}) = \mathbf{A}_n \mathbf{E} \mathbf{A}_n - \mathbf{B}_n \mathbf{E} \mathbf{B}_n.$$

Then $\mathcal{L}_n$ is invertible on the space of $p \times p$ matrices and $\|\mathcal{L}_n^{-1}\| = O_p(n^{-2})$. Consequently, if $\mathcal{L}_n(\mathbf{E}_n) = O_p^m(1)$, then $\mathbf{E}_n = O_p^m(n^{-2})$.

Proof: Since $\mathbf{A}_n = \mathbf{B}_n + \mathbf{C}_n$,

$$\mathcal{L}_n(\mathbf{E}) = \mathbf{A}_n \mathbf{E} \mathbf{A}_n - \mathbf{B}_n \mathbf{E} \mathbf{B}_n = \mathbf{C}_n \mathbf{E} \mathbf{A}_n + \mathbf{B}_n \mathbf{E} \mathbf{C}_n.$$

Vectorizing gives $\text{vec}\{\mathcal{L}_n(\mathbf{E})\} = (\mathbf{A}_n \otimes \mathbf{C}_n + \mathbf{C}_n \otimes \mathbf{B}_n) \text{vec}(\mathbf{E})$, where we have used symmetry of $\mathbf{A}_n$, $\mathbf{B}_n$, and $\mathbf{C}_n$.

Let $\mathbf{K}_n = \mathbf{A}_n \otimes \mathbf{C}_n + \mathbf{C}_n \otimes \mathbf{B}_n$. Since $\mathbf{A}_n$ and $\mathbf{C}_n$ are positive definite and $\mathbf{B}_n$ is positive semidefinite, $\mathbf{K}_n$ is positive definite. Moreover,

$$\lambda_{\min}(\mathbf{K}_n) \geq \lambda_{\min}(\mathbf{A}_n \otimes \mathbf{C}_n) = \lambda_{\min}(\mathbf{A}_n)\lambda_{\min}(\mathbf{C}_n).$$

By assumption, $\lambda_{\min}(\mathbf{A}_n)$ and $\lambda_{\min}(\mathbf{C}_n)$ are bounded below by a positive multiple of n , with probability tending to one. Hence, $\lambda_{\min}(\mathbf{K}_n) \geq \lambda_{\min}(\mathbf{A}_n)\lambda_{\min}(\mathbf{C}_n)$ is bounded below by a positive multiple of n^2 , with probability tending to one. Therefore,

$$\|\mathbf{K}_n^{-1}\| = O_p(n^{-2}).$$

Thus, $\|\mathcal{L}_n^{-1}\| = O_p(n^{-2})$, and so if $\mathcal{L}_n(\mathbf{E}_n) = O_p^m(1)$, then $\text{vec}(\mathbf{E}_n) = \mathbf{K}_n^{-1} \text{vec}(\mathcal{L}_n(\mathbf{E}_n)) = O_p(n^{-2}) O_p(1)$, and since p is fixed, this implies $\mathbf{E}_n = O_p^m(n^{-2})$.

We are now ready to prove Theorem 5.

Proof of Theorem 5: We prove the result at the fixed point of Algorithm 8. Let

$$\mathbf{A} = \mathbf{Z}^T \mathbf{Z} + \mathbf{D}, \quad \text{so that} \quad \mathbf{S} = \mathbf{A}^{-1}.$$

Since p is fixed and by (A3) the design satisfies $n^{-1} \mathbf{Z}^T \mathbf{Z} \rightarrow \boldsymbol{\Sigma}_x \in \mathcal{S}_p^+$. Thus, $\mathbf{A} = O_p^m(n)$ and $\mathbf{S} = O_p^m(n^{-1})$. By (A4), at the fixed point, $\tilde{\boldsymbol{\Sigma}} = O_p^m(n^{-1})$. The weighted empirical sums appearing below are controlled by the empirical order conditions in (A5). These conditions hold, for example, under standard bounded design conditions, or under random-design conditions with sufficiently many moments.

The posterior mode $\hat{\boldsymbol{\beta}}$ satisfies

$$\mathbf{Z}^T \zeta_1(\mathbf{Z}\hat{\boldsymbol{\beta}}) - \mathbf{D}\hat{\boldsymbol{\beta}} = \mathbf{0} \quad \text{or equivalently} \quad \hat{\boldsymbol{\beta}} = \mathbf{S}\mathbf{Z}^T[\mathbf{Z}\hat{\boldsymbol{\beta}} + \zeta_1(\mathbf{Z}\hat{\boldsymbol{\beta}})].$$

Let $\mathbf{s}^2 = \text{dg}(\mathbf{Z}\tilde{\boldsymbol{\Sigma}}\mathbf{Z}^T)$. Since $\tilde{\boldsymbol{\Sigma}} = O_p^m(n^{-1})$, we have $s_i^2 = O_p(n^{-1}\|\mathbf{z}_i\|^2)$. By a Taylor series

expansion of ξ_1 around $s_i^2 = 0$,

$$\xi_1(\mathbf{Z}\tilde{\boldsymbol{\mu}}, \mathbf{s}^2) = \zeta_1(\mathbf{Z}\tilde{\boldsymbol{\mu}}) + \boldsymbol{\Delta}_1,$$

where $\mathbf{Z}^T \boldsymbol{\Delta}_1 = O_p^m(1)$ by (A5). At the fixed point of Algorithm 8 we have $\tilde{\boldsymbol{\mu}} = \mathbf{S}\mathbf{Z}^T[\mathbf{Z}\tilde{\boldsymbol{\mu}} + \xi_1(\mathbf{Z}\tilde{\boldsymbol{\mu}}, \mathbf{s}^2)]$. Subtracting the corresponding equation for $\hat{\boldsymbol{\beta}}$ gives

$$\mathbf{A}(\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\beta}}) = \mathbf{Z}^T \mathbf{Z}(\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\beta}}) + \mathbf{Z}^T[\zeta_1(\mathbf{Z}\tilde{\boldsymbol{\mu}}) - \zeta_1(\mathbf{Z}\hat{\boldsymbol{\beta}})] + \mathbf{Z}^T \boldsymbol{\Delta}_1.$$

By the mean value theorem, $\zeta_1(\mathbf{Z}\tilde{\boldsymbol{\mu}}) - \zeta_1(\mathbf{Z}\hat{\boldsymbol{\beta}}) = \mathbf{W}\mathbf{Z}(\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\beta}})$, where $\mathbf{W}$ is diagonal with entries between $\zeta_2(\mathbf{z}_i^T \tilde{\boldsymbol{\mu}})$ and $\zeta_2(\mathbf{z}_i^T \hat{\boldsymbol{\beta}})$. Hence,

$$(\mathbf{D} - \mathbf{Z}^T \mathbf{W}\mathbf{Z})(\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\beta}}) = \mathbf{Z}^T \boldsymbol{\Delta}_1.$$

Since $-\zeta_2(x) > 0$ for all x the matrix $\mathbf{D} - \mathbf{Z}^T \mathbf{W}\mathbf{Z}$ has eigenvalues of order n under the regularity conditions above, in particular (A6). Therefore, $\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\beta}} = O_p^m(n^{-1})$.

We now turn to the covariance. At the fixed point, define

$$\begin{aligned} \mathbf{W}_\zeta &= \text{diag}[\zeta_2(\mathbf{Z}\tilde{\boldsymbol{\mu}})], \quad \mathbf{B}_\zeta = \mathbf{Z}^T(\mathbf{I}_n + \mathbf{W}_\zeta)\mathbf{Z}, \quad \mathbf{C}_\zeta = \mathbf{A} - \mathbf{B}_\zeta = \mathbf{D} - \mathbf{Z}^T \mathbf{W}_\zeta \mathbf{Z} \quad \text{and} \\ \mathbf{W}_\xi &= \text{diag}[\xi_2(\mathbf{Z}\tilde{\boldsymbol{\mu}}, \mathbf{s}^2)], \quad \mathbf{B}_\xi = \mathbf{Z}^T(\mathbf{I}_n + \mathbf{W}_\xi)\mathbf{Z}, \quad \mathbf{C}_\xi = \mathbf{A} - \mathbf{B}_\xi = \mathbf{D} - \mathbf{Z}^T \mathbf{W}_\xi \mathbf{Z}. \end{aligned}$$

Note that the matrix $\mathbf{B}_\zeta$ is positive semidefinite because $1 + \zeta_2(x) > 0$ for all x (this is the variance of a unit-variance normal variable truncated below at zero).

Using another Taylor series expansion in s_i^2 we have

$$\xi_2(\mathbf{Z}\tilde{\boldsymbol{\mu}}, \mathbf{s}^2) = \zeta_2(\mathbf{Z}\tilde{\boldsymbol{\mu}}) + \boldsymbol{\Delta}_2,$$

where $\mathbf{Z}^T \text{diag}(\boldsymbol{\Delta}_2)\mathbf{Z} = O_p^m(1)$. Thus, $\mathbf{B}_\xi = \mathbf{B}_\zeta + O_p^m(1)$. The covariance fixed-point equation in Algorithm 8 can be rewritten as

$$\tilde{\boldsymbol{\Sigma}} = \mathbf{S} + \mathbf{S}\mathbf{B}_\xi\mathbf{S} + \mathbf{S}\mathbf{B}_\zeta\tilde{\boldsymbol{\Sigma}}\mathbf{B}_\zeta\mathbf{S}.$$

Multiplying on the left and right by $\mathbf{A} = \mathbf{S}^{-1}$, using $\mathbf{B}_\xi = \mathbf{B}_\zeta + O_p^m(1)$, and rearranging gives

$$\mathbf{A}\tilde{\Sigma}\mathbf{A} - \mathbf{B}_\zeta\tilde{\Sigma}\mathbf{B}_\zeta = \mathbf{A} + \mathbf{B}_\zeta + O_p^m(1).$$

Now, since $\mathbf{C}_\zeta = \mathbf{A} - \mathbf{B}_\zeta$ direct expansion shows that $\mathbf{A}\mathbf{C}_\zeta^{-1}\mathbf{A} - \mathbf{B}_\zeta\mathbf{C}_\zeta^{-1}\mathbf{B}_\zeta = \mathbf{A} + \mathbf{B}_\zeta$. Also, since $\mathbf{A} = \mathbf{B}_\zeta + \mathbf{C}_\zeta$, $(\mathbf{B}_\zeta + \mathbf{C}_\zeta)\mathbf{C}_\zeta^{-1}(\mathbf{B}_\zeta + \mathbf{C}_\zeta) - \mathbf{B}_\zeta\mathbf{C}_\zeta^{-1}\mathbf{B}_\zeta = \mathbf{B}_\zeta + \mathbf{C}_\zeta + \mathbf{B}_\zeta = \mathbf{A} + \mathbf{B}_\zeta$. Therefore, if $\mathbf{E} = \tilde{\Sigma} - \mathbf{C}_\zeta^{-1}$, then $\mathbf{A}\mathbf{E}\mathbf{A} - \mathbf{B}_\zeta\mathbf{E}\mathbf{B}_\zeta = O_p^m(1)$. By Lemma 4, the associated linear operator $\mathcal{L}_n(\mathbf{E}) = \mathbf{A}\mathbf{E}\mathbf{A} - \mathbf{B}_\zeta\mathbf{E}\mathbf{B}_\zeta$ has inverse of order $O_p(n^{-2})$, since $\mathbf{A} = \mathbf{B}_\zeta + \mathbf{C}_\zeta$, $\mathbf{B}_\zeta$ is positive semidefinite, and $\mathbf{A}$ and $\mathbf{C}_\zeta$ are positive definite with eigenvalues of order n . Hence, $\mathbf{E} = O_p^m(n^{-2})$, and therefore $\tilde{\Sigma} = \mathbf{C}_\zeta^{-1} + O_p^m(n^{-2})$.

It remains to replace $\mathbf{C}_\zeta$ by the curvature matrix evaluated at the posterior mode. Since, $\tilde{\boldsymbol{\mu}} - \hat{\boldsymbol{\beta}} = O_p^m(n^{-1})$, another Taylor series expansion gives

$$\mathbf{Z}^T \left[\text{diag}\{\zeta_2(\mathbf{Z}\tilde{\boldsymbol{\mu}})\} - \text{diag}\{\zeta_2(\mathbf{Z}\hat{\boldsymbol{\beta}})\} \right] \mathbf{Z} = O_p^m(1).$$

This follows from the preceding mean bound under bounded-design conditions; for random designs it is included in the empirical order assumptions. Thus, $\mathbf{C}_\zeta = \mathbf{D} - \mathbf{Z}^T \text{diag}\{\zeta_2(\mathbf{Z}\hat{\boldsymbol{\beta}})\}\mathbf{Z} + O_p^m(1)$. Since the leading matrix has eigenvalues of order n , the inverse perturbation bound lemma above gives

$$\mathbf{C}_\zeta^{-1} = [\mathbf{D} - \mathbf{Z}^T \text{diag}\{\zeta_2(\mathbf{Z}\hat{\boldsymbol{\beta}})\}\mathbf{Z}]^{-1} + O_p^m(n^{-2}).$$

Combining this with the previous display leads to

$$\tilde{\Sigma} = [\mathbf{Z}^T \text{diag}\{-\zeta_2(\mathbf{Z}\hat{\boldsymbol{\beta}})\}\mathbf{Z} + \mathbf{D}]^{-1} + O_p^m(n^{-2}).$$

Since p is fixed, element-wise $O_p(n^{-2})$ convergence implies Frobenius convergence of the same order, i.e.,

$$\|\tilde{\Sigma} - \mathbf{V}\|_F = O_p(n^{-2}),$$

where $\mathbf{V} = [\mathbf{Z}^T \text{diag}\{-\zeta_2(\mathbf{Z}\hat{\boldsymbol{\beta}})\}\mathbf{Z} + \mathbf{D}]^{-1}$. Together with the mean result above, this proves the theorem.

H Additional numerical results for logistic regression

This section reports the logistic regression analogue of the probit regression results in Section 5.2. The same datasets, prior specification, performance metrics, and computational setup are used. The purpose of these results is to verify that the behaviour observed for probit regression is not specific to the probit likelihood.

Data	Method	RMSE(μ)	RMSE(σ)	MIN	Q1.	Q2.	Frob.	Time (s)
Breastfeed	stan	0.007	0.0108	97.6	97.9	98.1	0.016	0.493
	Laplace	0.258	0.0705	81.5	88.1	90.8	0.058	0.003
	MFVB	0.154	0.2600	77.6	78.9	84.5	0.185	0.002
	EP	0.008	0.0297	94.8	96.5	98.4	0.029	0.004
	GVB	0.006	0.0346	94.6	96.2	98.3	0.033	0.008
	GMP-L	0.038	0.0313	97.9	98.3	98.5	0.025	0.022
	GMP-G	0.004	0.0184	98.8	99.1	99.3	0.022	0.004
	GMP-E	0.006	0.0188	98.8	99.1	99.2	0.020	0.029
Ion.	stan	0.012	0.007	96.9	97.8	98.0	0.005	1.90
	Laplace	0.319	0.068	45.1	70.5	82.1	0.020	0.01
	MFVB	0.152	0.241	56.7	68.0	70.9	0.057	0.14
	EP	0.008	0.015	95.9	98.0	98.6	0.007	0.10
	GVB	0.012	0.035	94.3	96.6	97.0	0.012	0.11
	GMP-L	0.047	0.058	90.8	93.6	94.4	0.017	0.31
	GMP-G	0.012	0.033	95.2	96.9	97.3	0.011	0.04
	GMP-E	0.010	0.035	95.0	96.7	97.0	0.014	0.25
Spam	stan	0.007	0.004	97.2	97.8	98.0	0.0018	44.4
	Laplace	0.089	0.023	75.0	89.3	94.1	0.0039	0.3
	MFVB	0.200	0.225	11.8	57.1	71.2	0.0445	12.9
	EP	0.004	0.017	81.0	95.8	97.4	0.0035	4.4
	GVB	0.005	0.022	78.8	95.6	97.3	0.0045	4.4
	GMP-L	0.013	0.014	96.5	98.6	99.0	0.0046	4.1
	GMP-G	0.005	0.007	97.9	99.4	99.6	0.0026	0.8
	GMP-E	0.006	0.009	97.5	99.2	99.5	0.0025	5.1
Caravan	stan	0.015	0.016	96.9	97.8	98.1	0.025	1020.1
	Laplace	0.341	0.074	63.3	94.5	96.9	0.033	0.6
	MFVB	0.082	0.390	52.4	71.7	73.4	0.138	36.3
	EP	0.015	0.072	83.1	97.4	98.9	0.032	21.6
	GVB	0.235	0.159	69.9	94.5	97.8	0.072	21.5
	GMP-L	0.100	0.073	89.4	96.6	97.7	0.034	15.6
	GMP-G	0.138	0.110	81.1	96.5	98.2	0.054	39.4
	GMP-E	0.041	0.059	91.2	98.6	99.1	0.036	62.2

Table 5: Performance metrics for Logistic regression.

Table 5 shows broadly similar behaviour to the probit results in Table 2 of the main text. MFVB tends to underestimate posterior uncertainty and gives substantially lower marginal density accuracies on several datasets. EP and the GMP variants are generally the most accurate

methods. In particular, GMP often improves the marginal density accuracy of the approximation used for initialization, while remaining much faster than the short-run MCMC benchmark.

As in the probit case, the quality of GMP depends on the initialization, and no single initialization dominates across all datasets and metrics.

Unlike the probit results in the main text, there is no auxiliary-variable MP entry for logistic regression; the MP-based entries here are the GMP variants. We leave such a variant for future work.

I Bayesian Lasso details

This appendix gives the MFVB and GVB approximations used to initialize the GMP updates for the Bayesian Lasso example in Sections 4.6 and 5.3.

I.1 Model and latent-variable representation

We consider the model $y_i \mid \beta, \sigma^2 \sim N(\mathbf{x}_i^T \beta, \sigma^2)$ with

$$\beta_j \mid \sigma^2 \sim \text{Laplace}\left(0, \frac{\sigma}{\lambda}\right) \quad \text{that is} \quad p(\beta_j \mid \sigma^2) = \frac{\lambda}{2\sigma} \exp\left(-\frac{\lambda}{\sigma} |\beta_j|\right). \quad (76)$$

We treat $\lambda^2 > 0$ as fixed to simplify the analysis. The parameters of interest are $\theta = (\beta, \sigma^2)$. We assume that $\mathbf{y}$ and $\mathbf{X}$ are centered so that, without loss of generality, the intercept can be ignored. Since (76) is difficult to handle directly, we instead adopt an auxiliary representation. This entails replacing (76) with

$$\beta_j \mid \sigma^2, a_j \sim N\left(0, \frac{\sigma^2}{a_j \lambda^2}\right), \quad a_j \stackrel{\text{iid}}{\sim} \text{IG}(1, 1/2), \quad 1 \leq j \leq p, \quad (77)$$

Park and Casella (2008) show that we need to condition β_j on σ^2 for the regression coefficient posteriors to be unimodal. The above is equivalent to Park and Casella (2008), where we have shifted the dependence of a_j on λ^2 to having β_j depending on λ^2 . Finally, we use $\sigma^2 \sim \text{IG}(A, B)$ as the (conjugate) prior for σ^2 .

I.2 Mean Field Variational Bayes

The VB approximation corresponding to the factorization $q(\beta, \sigma^2, \mathbf{a}) = q(\beta)q(\sigma^2)q(\mathbf{a})$ leads to

$$\begin{aligned} q(\beta) &= N\left(\boldsymbol{\mu} \equiv \mathbf{Q}\mathbf{X}^T \mathbf{y}, \boldsymbol{\Sigma} \equiv (\tilde{B}/\tilde{A})\mathbf{Q}\right) \quad \text{and} \\ q(\sigma^2) &= \text{IG}\left(\tilde{A} \equiv A + \frac{n+p}{2}, \tilde{B} \equiv B + \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|^2 + \frac{\lambda^2}{2}\boldsymbol{\mu}^T \mathbf{D}\boldsymbol{\mu} + \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}(\mathbf{X}^T \mathbf{X} + \lambda^2 \mathbf{D})\right)\right). \end{aligned}$$

Since $p(\mathbf{a}|\mathcal{D}, \boldsymbol{\beta}, \sigma^2) = \prod_{j=1}^p p(a_j|\mathcal{D}, \beta_j, \sigma^2)$ we have $q(\mathbf{a}) = \prod_{j=1}^p$ and so we have independently

$$q(a_j) = \text{Inverse-Gaussian} \left(d_j \equiv \sqrt{\frac{\tilde{B}/\tilde{A}}{\lambda^2(\mu_j^2 + \Sigma_{jj}^2)}}, \lambda_j \equiv 1 \right)$$

where $\mathbf{Q} = (\mathbf{X}^T \mathbf{X} + \lambda^2 \mathbf{D})^{-1}$, $\mathbf{D} = \text{diag}(\mathbb{E}_q(\mathbf{a})) = \text{diag}(\mathbf{d})$. Here we use the inverse-Gaussian parameterization with density proportional to

$$x^{-3/2} \exp \left\{ -\frac{\lambda(x - m)^2}{2m^2 x} \right\}, \quad x > 0,$$

where m is the mean and λ is the shape.

I.3 Gaussian Variational Bayes

Consider the model representation (76) and suppose that $q(\boldsymbol{\beta}, \sigma^2) = q(\boldsymbol{\beta})q(\sigma^2)$ with $q(\boldsymbol{\beta}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then (ignoring additive constants)

$$\log p(\mathbf{y}, \boldsymbol{\beta}, \sigma^2) = -\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 - \frac{\lambda}{\sigma} \|\boldsymbol{\beta}\|_1 - \left(A + \frac{n+p}{2} + 1 \right) \log(\sigma^2) - B\sigma^{-2}.$$

For fixed $q(\sigma^2)$, given $q(\boldsymbol{\beta}) = N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the optimal values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ maximize (ignoring additive constants)

$$\text{ELBO} = -\frac{1}{2}c_1 \left[\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|_2^2 + \text{tr}(\mathbf{X}^T \mathbf{X} \boldsymbol{\Sigma}) \right] - c_2 \mathbb{E}_q(\|\boldsymbol{\beta}\|_1) + \frac{1}{2} \log |\boldsymbol{\Sigma}|$$

were $c_1 = \mathbb{E}_q(\sigma^{-2})$ and $c_2 = \lambda \mathbb{E}_q(\sigma^{-1})$. For $\beta_j \sim N(\mu_j, \Sigma_{jj})$, let $s_j = \sqrt{\Sigma_{jj}}$, and $r_j = \mu_j/s_j$ then

$$\mathbb{E}_q|\beta_j| = \mu_j \{ \Phi(r_j) - \Phi(-r_j) \} + 2s_j \phi(r_j).$$

and

$$\frac{\partial \mathbb{E}_q|\beta_j|}{\partial \mu_j} = \Phi(r_j) - \Phi(-r_j).$$

Then

$$\begin{aligned}\frac{\partial \text{ELBO}}{\partial \boldsymbol{\mu}} &= c_1 \mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\mu}) - c_2 [\Phi(\mathbf{r}) - \Phi(-\mathbf{r})] \quad \text{and} \\ \frac{\partial^2 \text{ELBO}}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^T} &= -c_1 \mathbf{X}^T \mathbf{X} - c_2 \text{diag} \left[2 \text{dg}(\boldsymbol{\Sigma})^{-1/2} \odot \phi(\mathbf{r}) \right].\end{aligned}$$

We also have

$$\frac{\partial \text{ELBO}}{\partial \Sigma_{ij}} = \frac{1}{2} \text{tr} \left[\left\{ \boldsymbol{\Sigma}^{-1} - c_1 \mathbf{X}^T \mathbf{X} - c_2 \text{diag} \left(2 \text{dg}(\boldsymbol{\Sigma})^{-1/2} \odot \phi(\mathbf{r}) \right) \right\} \mathbf{E}_{ij} \right].$$

This leads to the updates

- $\boldsymbol{\Sigma} \leftarrow \left[c_1 \mathbf{X}^T \mathbf{X} + c_2 \text{diag} \left(2 \text{dg}(\boldsymbol{\Sigma})^{-1/2} \odot \phi(\mathbf{r}) \right) \right]^{-1}$
- $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \boldsymbol{\Sigma} \left[c_1 \mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\mu}) - c_2 \{ \Phi(\mathbf{r}) - \Phi(-\mathbf{r}) \} \right].$

These equations are solved by fixed-point iteration, using the current values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ on the right-hand side. We found that this fixed point iteration can be unstable and usually requires the algorithm to be initialized at good starting values. However, this fixed point approach is much faster than optimizing the ELBO directly with respect to $\boldsymbol{\Sigma}$.

The optimal $q(\sigma^2)$ is of the form

$$q(\sigma^2) = Z^{-1} \exp \left[-(\tilde{A} + 1) \log(\sigma^2) - \tilde{B} \sigma^{-1} - \tilde{C} \sigma^{-2} \right]$$

where $\tilde{A} = A + \frac{n+p}{2}$, $\tilde{B} = \lambda \cdot \mathbb{E}_q(\|\boldsymbol{\beta}\|_1)$, $\tilde{C} = B + \frac{1}{2} \mathbb{E}_{q(\boldsymbol{\beta})}(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2)$, and

$$Z = \int_0^\infty \exp \left[-(\tilde{A} + 1) \log(t) - \tilde{B} t^{-1/2} - \tilde{C} t^{-1} \right] dt$$

with $t = \sigma^2$. The q -density is not the usual inverse-Gamma distribution due to the σ^{-1} term appearing and needs to be handled numerically. The normalizing constant Z and the moments $\mathbb{E}_q(\sigma^{-2})$, $\mathbb{E}_q(\sigma^{-1})$, and $\mathbb{E}_q(\sigma^2)$ are evaluated numerically using a composite trapezoid rule.

Supplementary References

Bodnar, T. and Y. Okhrin (2008). “Properties of the singular, inverse and generalized inverse partitioned Wishart distributions”. In: *Journal of Multivariate Analysis* 99.10, pp. 2389–2405.

- Mathai, A. M. and S. B. Provost (1992). *Quadratic Forms in Random Variables: Theory and Applications*. Vol. 126. Statistics: Textbooks and Monographs. New York: Marcel Dekker. ISBN: 9780824786915.
- Ormerod, J. T. and M. P. Wand (2010). “Explaining variational approximations”. In: *The American Statistician* 64, pp. 140–153.
- Park, T. and G. Casella (2008). “The Bayesian Lasso”. In: *Journal of the American Statistical Association* 103.482, pp. 681–686.
- Press, S. J. (2005). *Applied multivariate analysis: using Bayesian and frequentist methods of inference*. Courier Corporation.
- Rong, J.-Y., Z.-F. Lu, and X.-Q. Liu (2012). “On Quadratic Forms of Multivariate t Distribution with Applications”. In: *Communications in Statistics - Theory and Methods* 41.2, pp. 300–308.
- Von Rosen, D. (1988). “Moments for the Inverted Wishart Distribution”. In: *Scandinavian Journal of Statistics* 15.2, pp. 97–109.