Statistical Science
2019, Vol. 34, No. 3, 472–485
https://doi.org/10.1214/19-STS712

Conditionally Conjugate Mean-Field Variational Bayes for Logistic Models

Daniele Durante and Tommaso Rigon

Abstract. Variational Bayes (VB) is a common strategy for approximate Bayesian inference, but simple methods are only available for specific classes of models including, in particular, representations having conditionally conjugate constructions within an exponential family. Models with logit components are an apparently notable exception to this class, due to the absence of conjugacy among the logistic likelihood and the Gaussian priors for the coefficients in the linear predictor. To facilitate approximate inference within this widely used class of models, Jaakkola and Jordan (*Stat. Comput.* **10** (2000) 25–37) proposed a simple variational approach which relies on a family of tangent quadratic lower bounds of the logistic log-likelihood, thus restoring conjugacy between these approximate bounds and the Gaussian priors. This strategy is still implemented successfully, but few attempts have been made to formally understand the reasons underlying its excellent performance. Following a review on VB for logistic models, we cover this gap by providing a formal connection between the above bound and a recent Pólya-gamma data augmentation for logistic regression. Such a result places the computational methods associated with the aforementioned bounds within the framework of variational inference for conditionally conjugate exponential family models, thereby allowing recent advances for this class to be inherited also by the methods relying on Jaakkola and Jordan (*Stat. Comput.* **10** (2000) 25–37).

Key words and phrases: EM, logistic regression, Pólya-gamma data augmentation, quadratic approximation, variational Bayes.

1. INTRODUCTION

The increasing availability of high-dimensional and massive datasets has motivated a wide interest in strategies for Bayesian learning of posterior distributions, beyond the classical MCMC methods (e.g., Gelfand and Smith, 1990). Indeed, sampling algorithms can face severe computational bottlenecks in complex statistical models, thus motivating alternative solutions based on scalable and efficient optimization of approximate posterior distributions. Notable methodologies within this class are the Laplace approximation (e.g., Bishop, 2006, Chapter 4.4), variational Bayes (e.g., Bishop, 2006, Chapter 10.1) and expectation propagation (e.g., Bishop, 2006, Chapter 10.7), with variational inference providing a standard choice in several fields, as discussed in recent reviews by Blei, Kucukelbir and McAuliffe (2017) and Ormerod and Wand (2010). Refer also to Jordan et al. (1999) for a seminal introduction of variational inference from a statistical perspective.

Adapting the notation in Blei, Kucukelbir and McAuliffe (2017), VB aims at obtaining a tractable approximation $q^*(\boldsymbol{\theta})$ for the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ of the random parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\intercal$, in the model having joint density $p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ and the observed data $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$, with $p(\boldsymbol{\theta})$ denoting the prior for $\boldsymbol{\theta}$. This optimization problem is formally stated by minimizing the Kullback–Leibler (KL) diver-

Daniele Durante is Assistant Professor of Statistics, Department of Decision Sciences and affiliate to the Bocconi Institute for Data Science and Analytics, Bocconi University, Via Roentgen 1, Milan, Italy (e-mail: daniele.durante@unibocconi.it). Tommaso Rigon is Ph.D. Student in Statistics, Department of Decision Sciences, Bocconi University, Via Roentgen 1, Milan, Italy (e-mail: tommaso.rigon@phd.unibocconi.it).

gence (Kullback and Leibler, 1951)

$$\begin{aligned}\mathrm{KL}\big[q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\mathbf{y})\big] &= \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\,\mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})\,p(\mathbf{y})}{p(\mathbf{y},\boldsymbol{\theta})}\,\mathrm{d}\boldsymbol{\theta},\end{aligned} \tag{1.1}$$

with respect to $q(\boldsymbol{\theta}) \in \mathcal{Q}$, where $\mathcal{Q}$ is a tractable, yet sufficiently flexible, class of approximating densities. As is clear from equation (1.1), the calculation of the KL divergence between $q(\boldsymbol{\theta})$ and the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ requires the evaluation of the normalizing constant $p(\mathbf{y})$. Due to this, the above minimization problem is equivalently stated as the maximization of the evidence lower bound (ELBO) function

$$\begin{aligned}\mathrm{ELBO}\big[q(\boldsymbol{\theta})\big] &= \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\,\mathrm{d}\boldsymbol{\theta} \\ &= -\mathrm{KL}\big[q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\mathbf{y})\big] + \log p(\mathbf{y}),\end{aligned} \tag{1.2}$$

which does not require the evaluation of $\log p(\mathbf{y})$. In fact, since $\log p(\mathbf{y})$ does not depend on $\boldsymbol{\theta}$, maximizing (1.2) is equivalent to minimizing (1.1). Rewriting (1.2) as $\log p(\mathbf{y}) = \mathrm{ELBO}[q(\boldsymbol{\theta})] + \mathrm{KL}[q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\mathbf{y})]$ it can be additionally noticed that the ELBO provides a lower bound of $\log p(\mathbf{y})$ for every $q(\boldsymbol{\theta}) \in \mathcal{Q}$, since the Kullback–Leibler divergence in (1.1) is always nonnegative (Kullback and Leibler, 1951).

The above setup defines the general rationale underlying VB. However, recalling equation (1.2), the practical feasibility of the variational optimization requires a tractable form for the joint density $p(\mathbf{y},\boldsymbol{\theta})$ along with a simple, yet flexible, approximating family $\mathcal{Q}$. This is the case of mean-field VB for conditionally conjugate exponential family models having global and local variables (Wang and Titterington, 2004, Bishop, 2006, Hoffman et al., 2013, Blei, Kucukelbir and McAuliffe, 2017). Letting $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{z})$ and $\mathcal{Q} = \{q(\boldsymbol{\beta},\mathbf{z}) : q(\boldsymbol{\beta},\mathbf{z}) = q(\boldsymbol{\beta})\prod_{i=1}^n q(z_i)\}$, these methods focus on obtaining a mean-field approximation

$$\begin{aligned}q^*(\boldsymbol{\beta},\mathbf{z}) &= \operatorname*{argmin}_{q(\boldsymbol{\beta},\mathbf{z})\in\mathcal{Q}} \big\{\mathrm{KL}\big[q(\boldsymbol{\beta},\mathbf{z})\|p(\boldsymbol{\beta},\mathbf{z}|\mathbf{y})\big]\big\} \\ &= \operatorname*{argmax}_{q(\boldsymbol{\beta},\mathbf{z})\in\mathcal{Q}} \big\{\mathrm{ELBO}\big[q(\boldsymbol{\beta},\mathbf{z})\big]\big\},\end{aligned} \tag{1.3}$$

for the joint posterior density $p(\boldsymbol{\beta},\mathbf{z}|\mathbf{y})$ of the global parameters $\boldsymbol{\beta} = (\beta_1,\dots,\beta_p)^\intercal$ and the local variables $\mathbf{z} = (z_1,\dots,z_n)^\intercal$ in the model having joint density

$$p(\mathbf{y},\boldsymbol{\beta},\mathbf{z}) = p(\boldsymbol{\beta})\prod_{i=1}^n p(y_i,z_i|\boldsymbol{\beta}), \tag{1.4}$$

where $p(y_i,z_i|\boldsymbol{\beta}) = p(z_i|\boldsymbol{\beta})p(y_i|z_i,\boldsymbol{\beta})$ is from an exponential family and $p(\boldsymbol{\beta})$ defines a conjugate prior for such a density. The latent quantities $\mathbf{z}$, when present, typically denote random effects or unit-specific augmented data within a hierarchical formulation, such as in mixture models.

Although the above assumptions appear restrictive, the factorization of $q(\boldsymbol{\beta},\mathbf{z})$—characterizing the mean-field variational family $\mathcal{Q}$—provides a flexible class in several applications and also allows direct implementation of simple coordinate ascent variational inference (CAVI) routines (Bishop, 2006, Chapter 10.1.1) which sequentially maximize the ELBO in (1.3) with respect to each factor in $q(\boldsymbol{\beta},\mathbf{z}) = q(\boldsymbol{\beta})\prod_{i=1}^n q(z_i)$—fixing the others at their most recent update. Moreover, the exponential family and the conjugacy assumptions further simplify calculations by providing approximating densities $q^*(\boldsymbol{\beta})$ and $q^*(z_i)$, $i = 1,\dots,n$ from tractable classes of random variables. These advantages have further motivated some recent computational improvements (Hoffman et al., 2013) and theoretical studies (Wang and Titterington, 2004). We refer to Hoffman et al. (2013) and Blei, Kucukelbir and McAuliffe (2017) for details on the methods related to the general formulation in (1.3)–(1.4), and we focus here on models having logistic likelihoods as building blocks. Indeed, although the conjugacy and the exponential family assumptions are common to a variety of machine learning representations (e.g., Blei, Ng and Jordan, 2003; Airoldi et al., 2008; Hoffman et al., 2013) classical Bayesian logistic regression models

$$\begin{aligned}(y_i|\boldsymbol{\beta}) &\sim \mathrm{Bern}\left[\frac{\exp(\mathbf{x}_i^\intercal\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\intercal\boldsymbol{\beta})}\right], \quad i = 1,\dots,n, \\ \boldsymbol{\beta} &\sim \mathrm{N}_p(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0),\end{aligned} \tag{1.5}$$

do not enjoy direct conjugacy between the likelihood for the binary responses $y_i \in \{0,1\}$ and the Gaussian prior for the coefficients $\boldsymbol{\beta}$ in the linear predictor (e.g., Wang and Blei, 2013). This apparently notable exception to conditionally conjugate exponential family models also holds, as a direct consequence, for a wide set of formulations which incorporate Bayesian logistic regressions at some layer of the hierarchical specification. Some relevant examples are classification via Gaussian processes (Rasmussen and Williams, 2006), supervised nonparametric clustering (Ren et al., 2011) and hierarchical mixture of experts (Bishop and Svensén, 2003).

To implement tractable VB for nonconjugate models, several routines beyond conjugate mean-field VB

have been proposed (e.g., Jaakkola and Jordan, 2000; Braun and McAuliffe, 2010; Wand et al., 2011, Wang and Blei, 2013). In the context of logistic regression, Jaakkola and Jordan (2000) developed a seminal VB algorithm based on the quadratic lower bound

$$\begin{aligned}\log \bar{p}(y_i|\boldsymbol{\beta}) &= (y_i - 0.5)\mathbf{x}_i^\intercal \boldsymbol{\beta} - 0.5\xi_i \\ &\quad - 0.25\xi_i^{-1}\tanh(0.5\xi_i)\big[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 \\ &\quad - \xi_i^2\big] - \log[1+\exp(-\xi_i)],\end{aligned} \tag{1.6}$$

for the log-likelihood $\log p(y_i|\boldsymbol{\beta}) = y_i(\mathbf{x}_i^\intercal\boldsymbol{\beta}) - \log[1+\exp(\mathbf{x}_i^\intercal\boldsymbol{\beta})] \geq \log \bar{p}(y_i|\boldsymbol{\beta})$ of every $y_i \in \{0,1\}$ from a logistic regression. In equation (1.6), $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})^\intercal$ comprises the covariates measured for the i-th unit, whereas $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^\intercal$ denote the associated coefficients. The vector $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_n)^\intercal$ represents instead unit-specific variational parameters defining the location where $\log \bar{p}(y_i|\boldsymbol{\beta})$ is tangent to $\log p(y_i|\boldsymbol{\beta})$. In fact, $\log \bar{p}(y_i|\boldsymbol{\beta}) = \log p(y_i|\boldsymbol{\beta})$ when $\xi_i^2 = (\mathbf{x}_i^\intercal\boldsymbol{\beta})^2$.

Using equation (1.6), Jaakkola and Jordan (2000) developed an expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) in order to approximate the posterior density $p(\boldsymbol{\beta}|\mathbf{y})$ of $\boldsymbol{\beta}$. At the generic iteration t, such a routine alternates between an E-step in which the conditional density of the random coefficients $\boldsymbol{\beta}$—given the current value $\boldsymbol{\xi}^{(t-1)}$—is updated to obtain $q^{(t)}(\boldsymbol{\beta})$, and an M-step which calculates the expectation of the augmented approximate log-likelihood $\log \bar{p}(\mathbf{y},\boldsymbol{\beta}) = \log p(\boldsymbol{\beta}) + \sum_{i=1}^n \log \bar{p}(y_i|\boldsymbol{\beta})$ with respect to $q^{(t)}(\boldsymbol{\beta})$ and then maximizes it as a function of $\boldsymbol{\xi}$. Recalling the general presentation of EM by Bishop (2006) in Chapter 9.4, and Appendices A–B in Jaakkola and Jordan (2000), this strategy ultimately maximizes $\log \bar{p}(\mathbf{y}) = \log \int_{\Re^p} p(\boldsymbol{\beta}) \prod_{i=1}^n \bar{p}(y_i|\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta}$ with respect to $\boldsymbol{\xi}$, by sequentially optimizing the lower bound

$$\int_{\Re^p} q(\boldsymbol{\beta}) \log \frac{p(\boldsymbol{\beta})\prod_{i=1}^n \bar{p}(y_i|\boldsymbol{\beta})}{q(\boldsymbol{\beta})}\,\mathrm{d}\boldsymbol{\beta}, \tag{1.7}$$

as a function of the unknown density $q(\boldsymbol{\beta})$ and of the fixed parameters $\boldsymbol{\xi}$, where $p(\boldsymbol{\beta})$ defines the density of the Gaussian prior for $\boldsymbol{\beta}$. Hence, as is clear from Algorithm 1, this EM routine produces an optimal estimate $\boldsymbol{\xi}^*$ of $\boldsymbol{\xi}$ and, as a direct byproduct, also a density $q^*(\boldsymbol{\beta})$, which is regarded as an approximate posterior in Jaakkola and Jordan (2000). Indeed, recalling the EM structure, $q^*(\boldsymbol{\beta})$ coincides with the conditional density $\bar{p}^*(\boldsymbol{\beta}|\mathbf{y})$ obtained by updating the prior $p(\boldsymbol{\beta})$ with the approximate likelihood $\prod_{i=1}^n \bar{p}^*(y_i|\boldsymbol{\beta})$ induced by (1.6) and evaluated at the optimal variational parameters $\boldsymbol{\xi}^* = (\xi_1^*,\ldots,\xi_n^*)^\intercal$.

However, although this routine is successfully implemented in the machine learning and statistical literature (e.g., Bishop and Svensén, 2003; Rasmussen and

Algorithm 1: EM algorithm for approximate Bayesian inference by Jaakkola and Jordan (2000)

Initialize $\xi_1^{(0)},\ldots,\xi_n^{(0)}$.

for $t = 1$ until convergence of (1.7) **do**

Expectation. Update $q^{(t)}(\boldsymbol{\beta}) = \bar{p}^{(t-1)}(\boldsymbol{\beta}|\mathbf{y}) \propto p(\boldsymbol{\beta})\prod_{i=1}^n \bar{p}^{(t-1)}(y_i|\boldsymbol{\beta})$ to obtain a $\mathrm{N}_p(\boldsymbol{\mu}^{(t)},\boldsymbol{\Sigma}^{(t)})$ density with

$$\boldsymbol{\Sigma}^{(t)} = \big(\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^\intercal\bar{\mathbf{Z}}^{(t-1)}\mathbf{X}\big)^{-1}, \qquad \boldsymbol{\mu}^{(t)} = \boldsymbol{\Sigma}^{(t)}\big[\mathbf{X}^\intercal(\mathbf{y} - 0.5\cdot\mathbf{1}_n) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\big],$$

where $\bar{\mathbf{Z}}^{(t-1)} = \mathrm{diag}[0.5(\xi_1^{(t-1)})^{-1}\tanh(0.5\xi_1^{(t-1)}),\ldots,0.5(\xi_n^{(t-1)})^{-1}\tanh(0.5\xi_n^{(t-1)})]$ and $\mathbf{1}_n = (1,\ldots,1)^\intercal$. Note that the quadratic form of (1.6) restores conjugacy between the Gaussian prior for $\boldsymbol{\beta}$ and the approximated likelihood. To clarify this result note that, for every ξ_i, $\bar{p}(y_i|\boldsymbol{\beta})$ is proportional to the kernel of a Gaussian variable with mean $\mathbf{x}_i^\intercal\boldsymbol{\beta}$ and variance $2\xi_i\tanh(0.5\xi_i)^{-1}$ for the transformed data $2\xi_i\tanh(0.5\xi_i)^{-1}(y_i - 0.5)$.

Maximization. Compute $\boldsymbol{\xi}^{(t)} = \mathrm{argmax}_{\boldsymbol{\xi}} \int_{\Re^p} q^{(t)}(\boldsymbol{\beta})\log \bar{p}(\mathbf{y},\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta}$ to obtain the solutions

$$\xi_i^{(t)} = \big\{\mathbb{E}_{q^{(t)}(\boldsymbol{\beta})}\big[(\mathbf{x}_i^\intercal\boldsymbol{\beta})^2\big]\big\}^{1/2} = \big[\mathbf{x}_i^\intercal\boldsymbol{\Sigma}^{(t)}\mathbf{x}_i + (\mathbf{x}_i^\intercal\boldsymbol{\mu}^{(t)})^2\big]^{1/2}, \quad \text{for every } i = 1,\ldots,n.$$

Note that $\int_{\Re^p} q^{(t)}(\boldsymbol{\beta})\log \bar{p}(\mathbf{y},\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta} = \mathrm{const} + \sum_{i=1}^n \int_{\Re^p} q^{(t)}(\boldsymbol{\beta})\log \bar{p}(y_i|\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta}$. Hence, it is possible to maximize the expected log-likelihood associated with every y_i separately, as a function of each ξ_i, for $i = 1,\ldots,n$. This result leads to the above solution.

Output of the algorithm: $\boldsymbol{\xi}^*$ and, as a byproduct, the approximate posterior $q^*(\boldsymbol{\beta}) = \bar{p}^*(\boldsymbol{\beta}|\mathbf{y})$.

Williams, 2006; Lee, Huang and Hu, 2010; Ren et al., 2011; Carbonetto and Stephens, 2012; Tang, Browne and McNicholas, 2015; Wand, 2017), it is still not clear how the solution $q^*(\boldsymbol{\beta})$ relates to the formal VB setup in (1.1)–(1.2). Indeed, $\bar{p}^*(\boldsymbol{\beta}|\mathbf{y})$ is not the posterior induced by a Bayesian logistic regression. This is due to the fact that every $p(y_i|\boldsymbol{\beta})$ in the kernel of $p(\boldsymbol{\beta}|\mathbf{y})$ is replaced with the approximate likelihood $\bar{p}^*(y_i|\boldsymbol{\beta})$ evaluated at the optimal variational parameters $\boldsymbol{\xi}^*$ maximizing $\log \bar{p}(\mathbf{y})$. This last result, which is inherent to the EM (Dempster, Laird and Rubin, 1977), suggests an heuristic intuition for why $q^*(\boldsymbol{\beta})$ may still provide a reasonable approximation. Indeed, since $\log \bar{p}(y_i|\boldsymbol{\beta}) \leq \log p(y_i|\boldsymbol{\beta})$ for every ξ_i and $i = 1, \ldots, n$, the same holds for $\log \bar{p}(\mathbf{y})$ and $\log p(\mathbf{y})$. Thus, since $\log p(\mathbf{y})$ does not vary with $\boldsymbol{\xi}$, maximizing $\log \bar{p}(\mathbf{y})$ with respect to $\boldsymbol{\xi}$ is expected to provide the tightest approximation of each $\log p(y_i|\boldsymbol{\beta})$ via the lower bound in (1.6) evaluated at the optimum ξ_i^*, for $i = 1, \ldots, n$. This could guarantee similar predictive densities $p(\mathbf{y})$ and $\bar{p}^*(\mathbf{y})$. Hence, in correspondence to $\boldsymbol{\xi}^*$, the minimization of $\mathrm{KL}[q(\boldsymbol{\beta}) \| \bar{p}^*(\boldsymbol{\beta}|\mathbf{y})]$ in the E-step, would hopefully provide a solution $q^*(\boldsymbol{\beta}) = \bar{p}^*(\boldsymbol{\beta}|\mathbf{y})$ close to the true posterior $p(\boldsymbol{\beta}|\mathbf{y})$.

Although the above discussion provides an intuition for the excellent performance of the methods proposed by Jaakkola and Jordan (2000), it is important to notice that finding the tightest bound within a class of functions might not be sufficient if such a class is not flexible enough. Indeed, the quadratic form of (1.6) might be restrictive for logistic log-likelihoods, and therefore even the optimal approximation could fail to mimic $\log p(y_i|\boldsymbol{\beta})$. Moreover, according to (1.1), a formal VB setup requires the minimization of a well-defined KL divergence between an exact posterior and an approximating density from a specified variational family. Instead, Jaakkola and Jordan (2000) seem to minimize the divergence between an approximate posterior and a prespecified density. If this were the case, then their methods could be only regarded as an approximate solution to formal VB. Indeed, although (1.6) has been recently studied (de Leeuw and Lange, 2009; Browne and McNicholas, 2015), this is currently the main view of the EM summarized in Algorithm 1 (e.g., Blei, Kucukelbir and McAuliffe, 2017; Wang and Blei, 2013; Bishop, 2006).

In Section 2, we prove that this is not true and that (1.6), although apparently supported by purely mathematical arguments, has indeed a clear probabilistic interpretation related to a recent Pólya-gamma data augmentation for logistic regression (Polson, Scott and Windle, 2013). In particular, let $q(z_i)$ be the density of a Pólya-gamma $\mathrm{PG}(1, \xi_i)$, then (1.6) is a proper evidence lower bound associated with a VB approximation of the posterior for z_i in the conditional model $p(y_i, z_i|\boldsymbol{\beta})$ for data y_i from (1.5) and the Pólya-gamma variable $(z_i|\boldsymbol{\beta}) \sim \mathrm{PG}(1, \mathbf{x}_i^\intercal \boldsymbol{\beta})$, with $\boldsymbol{\beta}$ kept fixed. Combining such a result with the objective function defined in equation (1.7), allows us to formalize Algorithm 1 as a pure CAVI which approximates the joint posterior for $\boldsymbol{\beta}$ and the augmented Pólya-gamma data $z_1, \ldots, z_n$, under a mean-field variational approximation within the conditionally conjugate exponential family framework. These results are discussed in Section 3, and are further generalized to allow stochastic variational inference (Hoffman et al., 2013) in logistic models, thus covering an important computational gap. A final discussion can be found in Section 4. Codes and additional empirical assessments are available at https://github.com/tommasorigon/logisticVB. Although we focus on Bayesian inference, it shall be noticed that (1.6) motivates also an EM for maximum likelihood estimation of $\boldsymbol{\beta}$ (Jaakkola and Jordan, 2000). We study the properties of this EM in the Appendix.

2. CONDITIONALLY CONJUGATE VARIATIONAL REPRESENTATION

This section discusses the theoretical connection between equation (1.6) and a recent Pólya-gamma data augmentation for conditionally conjugate inference in Bayesian logistic models (Polson, Scott and Windle, 2013), thus allowing us to recast the methods proposed by Jaakkola and Jordan (2000) within the wider framework of mean-field variational inference for conditionally conjugate exponential families. We shall emphasize that, in a recent manuscript, Scott and Sun (2013) proposed an EM for maximum a posteriori estimation of $\boldsymbol{\beta}$ in (1.5), discussing connections with the variational strategies in Jaakkola and Jordan (2000). Their findings are however limited to computational differences and similarities between the two methods and the associated algorithms. We instead provide a fully probabilistic connection between the contribution by Jaakkola and Jordan (2000) and the results in Polson, Scott and Windle (2013), thus opening new avenues for advances in VB for logistic models.

To anticipate Lemma 2.1, note that the core contribution of Polson, Scott and Windle (2013) is in showing that $p(y_i|\boldsymbol{\beta})$ in model (1.5) can be expressed as a scale-mixture of Gaussians with respect to a Pólya-gamma density. This result facilitates the implementation of MCMC algorithms which update $\boldsymbol{\beta}$ and the

Pólya-gamma augmented data $\mathbf{z} = (z_1, \ldots, z_n)^\intercal$ from conjugate full conditionals. In fact, the joint density $p(\mathbf{y}, \mathbf{z}|\boldsymbol{\beta})$ has a Gaussian kernel in $\boldsymbol{\beta}$, thereby restoring Gaussian-Gaussian conjugacy in the full conditional. As discussed in Lemma 2.1, this data augmentation, although developed a decade later, was implicitly hidden in the bound of Jaakkola and Jordan (2000).

LEMMA 2.1. *Let* $\log \bar{p}(y_i|\boldsymbol{\beta})$ *denote the quadratic lower bound in* (1.6) *proposed by Jaakkola and Jordan* (2000) *for the logistic log-likelihood* $\log p(y_i|\boldsymbol{\beta})$ *of model* (1.5). *Then, for every* $i = 1, \ldots, n$, *we have*

$$\begin{aligned}\log \bar{p}(y_i|\boldsymbol{\beta}) &= \int_{\Re_+} q(z_i) \log \frac{p(y_i, z_i|\boldsymbol{\beta})}{q(z_i)} \,\mathrm{d}z_i \\ &= \mathbb{E}_{q(z_i)}\left[\log \frac{p(y_i, z_i|\boldsymbol{\beta})}{q(z_i)}\right],\end{aligned} \tag{2.1}$$

where $p(y_i, z_i|\boldsymbol{\beta}) = p(y_i|\boldsymbol{\beta})p(z_i|\boldsymbol{\beta})$ *and* $p(y_i|\boldsymbol{\beta}) = \exp(y_i \mathbf{x}_i^\intercal \boldsymbol{\beta})[1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})]^{-1}$, *while* $q(z_i)$ *and* $p(z_i|\boldsymbol{\beta})$ *are the densities of the Pólya-gamma random variables* $\mathrm{PG}(1, \xi_i)$ *and* $\mathrm{PG}(1, \mathbf{x}_i^\intercal \boldsymbol{\beta})$, *respectively.*

PROOF. To prove Lemma 2.1, first note that $0.5\xi_i + \log[1 + \exp(-\xi_i)] = \log[2\cosh(0.5\xi_i)]$ and $0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta}) = \log[1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})] - \log\{2\cosh[0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta})]\}$. Replacing these quantities in (1.6), we obtain

$$\begin{aligned}&y_i \mathbf{x}_i^\intercal \boldsymbol{\beta} - 0.25\xi_i^{-1} \tanh(0.5\xi_i)\left[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 - \xi_i^2\right] \\ &\quad + \log\{\cosh(0.5\xi_i)^{-1} \cosh[0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta})]\} \\ &\quad - \log[1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})].\end{aligned}$$

To highlight equation (2.1) in the above expression, first notice that $-0.25\xi_i^{-1} \tanh(0.5\xi_i)[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 - \xi_i^2] = \mathbb{E}_{q(z_i)}[-0.5 z_i (\mathbf{x}_i^\intercal \boldsymbol{\beta})^2] - \mathbb{E}_{q(z_i)}(-0.5 z_i \xi_i^2)$, where the expected value is taken with respect to $z_i \sim \mathrm{PG}(1, \xi_i)$ (Polson, Scott and Windle, 2013). Hence, $\log \bar{p}(y_i|\boldsymbol{\beta})$ can be expressed as

$$\begin{aligned}&\mathbb{E}_{q(z_i)}\left[\log\left[\frac{\exp(y_i \mathbf{x}_i^\intercal \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})}\right.\right. \\ &\quad \left.\left. \cdot \frac{\exp[-0.5 z_i (\mathbf{x}_i^\intercal \boldsymbol{\beta})^2] \cosh[0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta})] p(z_i)}{\exp(-0.5 z_i \xi_i^2) \cosh(0.5\xi_i) p(z_i)}\right]\right].\end{aligned}$$

Based on this expression, the proof is concluded after noticing that $\exp(y_i \mathbf{x}_i^\intercal \boldsymbol{\beta})[1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})]^{-1} = p(y_i|\boldsymbol{\beta})$, whereas $\exp[-0.5 z_i (\mathbf{x}_i^\intercal \boldsymbol{\beta})^2] \cosh[0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta})] p(z_i)$ and $\exp(-0.5 z_i \xi_i^2) \cosh(0.5\xi_i) p(z_i)$ characterize the densities $p(z_i|\boldsymbol{\beta})$ and $q(z_i)$ of the Pólya-gamma random variables $\mathrm{PG}(1, \mathbf{x}_i^\intercal \boldsymbol{\beta})$ and $\mathrm{PG}(1, \xi_i)$, respectively, with $p(z_i)$ the density of a $\mathrm{PG}(1, 0)$. □

According to Lemma 2.1, the expansion in equation (1.6) is a proper ELBO for $p(y_i|\boldsymbol{\beta})$ related to a VB approximation of the posterior for z_i in the conditional model $p(y_i, z_i|\boldsymbol{\beta})$ for the response data y_i from (1.5) and the local variable $(z_i|\boldsymbol{\beta}) \sim \mathrm{PG}(1, \mathbf{x}_i^\intercal \boldsymbol{\beta})$, with $\boldsymbol{\beta}$ kept fixed. Note that, although some intuition on the relation between $\log \bar{p}(y_i|\boldsymbol{\beta})$ and $\mathbb{E}_{q(z_i)}[\log p(y_i, z_i|\boldsymbol{\beta})]$ can be deduced from Scott and Sun (2013), the authors leave out additive constants not depending on $\boldsymbol{\beta}$ in $\log \bar{p}(y_i|\boldsymbol{\beta})$ when discussing this connection. According to Lemma 2.1, these quantities are crucial to formally interpret $\log \bar{p}(y_i|\boldsymbol{\beta})$ as a genuine ELBO, since they coincide with $-\mathbb{E}_{q(z_i)}[\log q(z_i)]$. Besides this result, Lemma 2.1 provides a formal characterization for the approximation error $\log p(y_i|\boldsymbol{\beta}) - \log \bar{p}(y_i|\boldsymbol{\beta})$. Indeed, adapting (1.2) to this setting, such a quantity is the KL divergence between a generic Pólya-gamma and the one obtained by conditioning on $\boldsymbol{\beta}$. This allows to complete $\log p(y_i|\boldsymbol{\beta}) \geq \log \bar{p}(y_i|\boldsymbol{\beta})$, as

$$\begin{aligned}&\log p(y_i|\boldsymbol{\beta}) \\ &\quad = \log \bar{p}(y_i|\boldsymbol{\beta}) + \mathrm{KL}\big[q(z_i) \| p(z_i|y_i, \boldsymbol{\beta})\big] \\ &\quad = \log \bar{p}(y_i|\boldsymbol{\beta}) + \mathrm{KL}\big[q(z_i) \| p(z_i|\boldsymbol{\beta})\big],\end{aligned} \tag{2.2}$$

where the last equality follows directly from the fact that $p(y_i, z_i|\boldsymbol{\beta}) = p(y_i|\boldsymbol{\beta})p(z_i|\boldsymbol{\beta})$ and, hence, $p(z_i|y_i, \boldsymbol{\beta}) = p(z_i|\boldsymbol{\beta})$. This result sheds light on the heuristic interpretation of $q^*(\boldsymbol{\beta})$ in Section 1. Indeed, as is clear from equation (2.2), if $q(z_i)$—evaluated at the optimal ξ_i^*—is globally close to $p(z_i|\boldsymbol{\beta})$ for every $\boldsymbol{\beta}$ and $i = 1, \ldots, n$, then (1.6) ensures accurate approximation of $\log p(y_i|\boldsymbol{\beta})$, thus providing approximate posteriors $q^*(\boldsymbol{\beta})$ close to $p(\boldsymbol{\beta}|\mathbf{y})$. Theorem 2.1 formalizes this result, proving that Algorithm 1 maximizes the ELBO of a well-defined model under a mean-field VB.

THEOREM 2.1. *The lower bound in equation* (1.7) *maximized by Jaakkola and Jordan* (2000) *in their* EM *for approximate Bayesian inference in model* (1.5) *coincides with the evidence lower bound* $\mathrm{ELBO}[q(\boldsymbol{\beta}, \mathbf{z})]$ *defined as*

$$\begin{aligned}&\int_{\Re^p \times \Re_+^n} q(\boldsymbol{\beta}, \mathbf{z}) \log \frac{p(\mathbf{y}, \boldsymbol{\beta}, \mathbf{z})}{q(\boldsymbol{\beta}, \mathbf{z})} \,\mathrm{d}\mathbf{z}\,\mathrm{d}\boldsymbol{\beta} \\ &\quad = \mathbb{E}_{q(\boldsymbol{\beta}, \mathbf{z})}\left[\log \frac{p(\mathbf{y}, \boldsymbol{\beta}, \mathbf{z})}{q(\boldsymbol{\beta}, \mathbf{z})}\right],\end{aligned} \tag{2.3}$$

where the joint density $p(\mathbf{y}, \boldsymbol{\beta}, \mathbf{z})$ *can be factorized as* $p(\boldsymbol{\beta}) \prod_{i=1}^n p(y_i, z_i|\boldsymbol{\beta}) = p(\boldsymbol{\beta}) \prod_{i=1}^n p(y_i|\boldsymbol{\beta}) p(z_i|\boldsymbol{\beta})$, *whereas* $q(\boldsymbol{\beta}, \mathbf{z}) = q(\boldsymbol{\beta}) \prod_{i=1}^n q(z_i)$, *with* $q(z_i)$ *and* $p(z_i|\boldsymbol{\beta})$ *denoting the densities of the two Pólya-gamma variables* $\mathrm{PG}(1, \xi_i)$ *and* $\mathrm{PG}(1, \mathbf{x}_i^\intercal \boldsymbol{\beta})$, *respectively.*

PROOF. The proof follows easily from Lemma 2.1. More specifically, let $\int_{\Re^p} q(\boldsymbol{\beta}) \sum_{i=1}^n \log \bar{p}(y_i|\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta} + \int_{\Re^p} q(\boldsymbol{\beta}) \log[p(\boldsymbol{\beta})q(\boldsymbol{\beta})^{-1}]\,\mathrm{d}\boldsymbol{\beta}$ denote an expanded representation of (1.7). Then, replacing $\log \bar{p}(y_i|\boldsymbol{\beta})$ with its probabilistic definition in (2.1) and performing simple mathematical calculations, we obtain

$$\sum_{i=1}^n \int_{\Re^p \times \Re_+} q(\boldsymbol{\beta})q(z_i) \log \frac{p(y_i, z_i|\boldsymbol{\beta})}{q(z_i)}\,\mathrm{d}z_i\,\mathrm{d}\boldsymbol{\beta} + \int_{\Re^p} q(\boldsymbol{\beta}) \log \frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})}\,\mathrm{d}\boldsymbol{\beta}.$$

Notice now that the second summand does not depend on $\mathbf{z}$, thereby allowing us to replace this integral with $\int_{\Re^p \times \Re_+^n} \log[p(\boldsymbol{\beta})q(\boldsymbol{\beta})^{-1}]q(\boldsymbol{\beta}) \prod_{i=1}^n q(z_i)\,\mathrm{d}\mathbf{z}\,\mathrm{d}\boldsymbol{\beta}$. Similar arguments can be made to incorporate $\prod_{i=1}^n q(z_i) = q(\mathbf{z})$ in the first integral. Making these substitutions in the above equation we obtain

$$\begin{aligned}
&\int_{\Re^p \times \Re_+^n} q(\boldsymbol{\beta})q(\mathbf{z}) \log \frac{p(\boldsymbol{\beta}) \prod_{i=1}^n p(y_i, z_i|\boldsymbol{\beta})}{q(\boldsymbol{\beta})q(\mathbf{z})}\,\mathrm{d}\mathbf{z}\,\mathrm{d}\boldsymbol{\beta} \\
&\quad = \int_{\Re^p \times \Re_+^n} q(\boldsymbol{\beta}, \mathbf{z}) \log \frac{p(\mathbf{y}, \boldsymbol{\beta}, \mathbf{z})}{q(\boldsymbol{\beta}, \mathbf{z})}\,\mathrm{d}\mathbf{z}\,\mathrm{d}\boldsymbol{\beta} \\
&\quad = \mathbb{E}_{q(\boldsymbol{\beta}, \mathbf{z})}\left[\log \frac{p(\mathbf{y}, \boldsymbol{\beta}, \mathbf{z})}{q(\boldsymbol{\beta}, \mathbf{z})}\right],
\end{aligned}$$

proving Theorem 2.1. Recall that $q(\boldsymbol{\beta}, \mathbf{z}) = q(\boldsymbol{\beta})q(\mathbf{z})$ with $q(\mathbf{z}) = \prod_{i=1}^n q(z_i)$ and $\int_{\Re_+} q(z_i)\,\mathrm{d}z_i = 1$. □

As clarified by Theorem 2.1, the methodology proposed by Jaakkola and Jordan (2000) coincides with a pure VB which minimizes the $\mathrm{KL}[q(\boldsymbol{\beta}, \mathbf{z}) \| p(\boldsymbol{\beta}, \mathbf{z}|\mathbf{y})]$ within the mean-field variational family $\mathcal{Q} = \{q(\boldsymbol{\beta}, \mathbf{z}) : q(\boldsymbol{\beta}, \mathbf{z}) = q(\boldsymbol{\beta}) \prod_{i=1}^n q(z_i)\}$ in the conditionally conjugate exponential family model having:

1. *Global variables*

$$\boldsymbol{\beta} \sim \mathrm{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \tag{2.4}$$

2. *Local variables*

$$(z_i|\boldsymbol{\beta}) \sim \mathrm{PG}\big(1, \mathbf{x}_i^\intercal \boldsymbol{\beta}\big), \tag{2.5}$$

independently for $i = 1, \ldots, n$.

3. *Binary response data*

$$(y_i|\boldsymbol{\beta}) \sim \mathrm{Bern}\left[\frac{\exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})}\right], \tag{2.6}$$

independently for $i = 1, \ldots, n$.

We refer to Section 2 in Choi and Hobert (2013) for this specific formulation of the Pólya-gamma data augmentation procedure which highlights how, unlike for the general specification in equation (1.4), the conditional distribution of each y_i does not depend on z_i. Recalling Section 1, this is not a necessary requirement. Indeed, what is important is that the joint density $p(y_i, z_i|\boldsymbol{\beta}) = p(y_i|\boldsymbol{\beta})p(z_i|\boldsymbol{\beta})$ of (2.5)–(2.6) is within an exponential family and the prior $p(\boldsymbol{\beta})$ is conjugate to it. Recalling Section 2 in Choi and Hobert (2013), and noticing that $\cosh[0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta})] = 0.5[1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})] \exp[-0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta})]$, this is the case of model (2.5)–(2.6). In fact

$$\begin{aligned}
&p(y_i, z_i|\boldsymbol{\beta}) \\
&\quad = \exp(y_i \mathbf{x}_i^\intercal \boldsymbol{\beta})[1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\beta})]^{-1} \\
&\qquad \cdot \exp\big[-0.5 z_i (\mathbf{x}_i^\intercal \boldsymbol{\beta})^2\big] \cosh\big[0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta})\big] p(z_i) \\
&\quad = 0.5 \exp\big[(y_i - 0.5)\mathbf{x}_i^\intercal \boldsymbol{\beta} \\
&\qquad - 0.5 z_i (\mathbf{x}_i^\intercal \boldsymbol{\beta})^2\big] p(z_i),
\end{aligned} \tag{2.7}$$

is proportional to the Gaussian kernel $\exp[(y_i - 0.5) \cdot \mathbf{x}_i^\intercal \boldsymbol{\beta} - 0.5 z_i (\mathbf{x}_i^\intercal \boldsymbol{\beta})^2]$, which is conjugate to the Gaussian prior for $\boldsymbol{\beta}$ in (2.4).

3. CAVI AND SVI FOR LOGISTIC MODELS

The results outlined in Section 2 recast the methods by Jaakkola and Jordan (2000) within a much broader framework motivating a formal CAVI and generalizations to stochastic variational inference (SVI).

3.1 Coordinate Ascent Variational Inference

As discussed in Section 1, the mean-field assumption allows the implementation of a simple CAVI algorithm (Blei, Kucukelbir and McAuliffe, 2017; Bishop, 2006, Chapter 10.1.1) which sequentially maximizes the evidence lower bound in (2.3) with respect to each factor in $q(\boldsymbol{\beta}) \prod_{i=1}^n q(z_i)$, via the following updates:

$$\begin{aligned}
q^{(t)}(\boldsymbol{\beta}) &= \frac{\exp\{\mathbb{E}_{q^{(t-1)}(\mathbf{z})}[\log p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{z})]\}}{c_{\boldsymbol{\beta}}(\mathbf{y})}, \\
q^{(t)}(z_i) &= \frac{\exp\{\mathbb{E}_{q^{(t)}(\boldsymbol{\beta})}[\log p(z_i|\mathbf{y}, \boldsymbol{\beta})]\}}{c_{z_i}(\mathbf{y})}, \\
&\qquad i = 1, \ldots, n,
\end{aligned} \tag{3.1}$$

at each iteration t—until convergence of the ELBO. In the above expressions, $c_{\boldsymbol{\beta}}(\mathbf{y})$ and $c_{z_i}(\mathbf{y})$, $i = 1, \ldots, n$, denote constants leading to proper densities. Note that in our case $p(z_i|\mathbf{y}, \boldsymbol{\beta}) = p(z_i|\mathbf{y}, \mathbf{z}_{-i}, \boldsymbol{\beta})$.

To clarify why (3.1) produces a routine which iteratively improves the ELBO, and ultimately maximizes it, note that, keeping fixed $q^{(t-1)}(z_1), \ldots, q^{(t-1)}(z_n)$,

equation (2.3) can be rewritten as

$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{\beta})}\left[\mathbb{E}_{q^{(t-1)}(\mathbf{z})}\left[\log\frac{p(\mathbf{y},\boldsymbol{\beta},\mathbf{z})}{q(\boldsymbol{\beta})p(\mathbf{y},\mathbf{z})}\right]\right]+\text{const}\\
&\quad=\mathbb{E}_{q(\boldsymbol{\beta})}\left[\mathbb{E}_{q^{(t-1)}(\mathbf{z})}\left[\log\frac{p(\boldsymbol{\beta}|\mathbf{y},\mathbf{z})}{q(\boldsymbol{\beta})}\right]\right]+\text{const}\\
&\quad=\mathbb{E}_{q(\boldsymbol{\beta})}\left[\log\frac{\exp\{\mathbb{E}_{q^{(t-1)}(\mathbf{z})}[\log p(\boldsymbol{\beta}|\mathbf{y},\mathbf{z})]\}}{q(\boldsymbol{\beta})c_{\boldsymbol{\beta}}(\mathbf{y})}\right]\\
&\qquad+\text{const},
\end{aligned}$$

where the last term is the negative KL among $q(\boldsymbol{\beta})$ and $\exp\{\mathbb{E}_{q^{(t-1)}(\mathbf{z})}[\log p(\boldsymbol{\beta}|\mathbf{y},\mathbf{z})]\}c_{\boldsymbol{\beta}}(\mathbf{y})^{-1}$, thus motivating the CAVI update for $q(\boldsymbol{\beta})$ in (3.1). Similar derivations can be done to obtain the solutions for $q(z_1),\dots,q(z_n)$ in (3.1). As is clear from (3.1), the CAVI solution identifies both the form of the approximating densities—without prespecifying them as part of the mean-field assumption—and the optimal parameters of these densities. As discussed in Section 1, these solutions are particularly easy to obtain under conditionally conjugate exponential family representations (Hoffman et al., 2013), including model (2.4)–(2.6). In fact, recalling Polson, Scott and Windle (2013), the full conditionals for the local and global variables in model (2.4)–(2.6) can be obtained via conditional conjugacy properties, which lead to

$$\begin{aligned}
&(\boldsymbol{\beta}|\mathbf{y},\mathbf{z})\sim \mathrm{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma}),\\
&(z_i|\mathbf{y},\boldsymbol{\beta})\sim \mathrm{PG}\big(1,\mathbf{x}_i^\intercal\boldsymbol{\beta}\big),\quad i=1,\dots,n,
\end{aligned}\tag{3.2}$$

where $\boldsymbol{\Sigma}=(\boldsymbol{\Sigma}_0^{-1}+\mathbf{X}^\intercal\mathbf{Z}\mathbf{X})^{-1}$ and $\boldsymbol{\mu}=\boldsymbol{\Sigma}[\mathbf{X}^\intercal(\mathbf{y}-0.5\cdot\mathbf{1}_n)+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0]$, with $\mathbf{Z}=\mathrm{diag}(z_1,\dots,z_n)$ and $\mathbf{X}$ the $n\times p$ design matrix with rows $\mathbf{x}_i^\intercal$, $i=1,\dots,n$. Moreover, recalling (3.2), both $(\boldsymbol{\beta}|\mathbf{y},\mathbf{z})$ and $(z_i|\mathbf{y},\boldsymbol{\beta})$ have the exponential family representation

$$\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y},\mathbf{z})&\propto\exp\{\boldsymbol{\eta}_1(\mathbf{y})^\intercal\boldsymbol{\beta}+\mathrm{vec}\big[\boldsymbol{\eta}_2(\mathbf{z})\big]^\intercal\\
&\qquad\cdot\mathrm{vec}\big(\boldsymbol{\beta}\boldsymbol{\beta}^\intercal\big)-\alpha\big[\boldsymbol{\eta}_1(\mathbf{y}),\boldsymbol{\eta}_2(\mathbf{z})\big]\},\\
p(z_i|\mathbf{y},\boldsymbol{\beta})&\propto\exp\big\{\eta_i(\boldsymbol{\beta})z_i-\alpha\big[\eta_i(\boldsymbol{\beta})\big]\big\}p(z_i),\\
&\qquad i=1,\dots,n,
\end{aligned}\tag{3.3}$$

with $\boldsymbol{\eta}_1(\mathbf{y})=\mathbf{X}^\intercal(\mathbf{y}-0.5\cdot\mathbf{1}_n)+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$, $\boldsymbol{\eta}_2(\mathbf{z})=-0.5(\boldsymbol{\Sigma}_0^{-1}+\mathbf{X}^\intercal\mathbf{Z}\mathbf{X})$ and $\eta_i(\boldsymbol{\beta})=-0.5(\mathbf{x}_i^\intercal\boldsymbol{\beta})^2$, denoting the natural parameters. Replacing these expressions in (3.1), it can be easily noticed that the CAVI solutions have the same density of the corresponding full conditionals with optimal natural parameters

$$\begin{aligned}
\boldsymbol{\lambda}_1^{(t)}&=\mathbb{E}_{q^{(t-1)}(\mathbf{z})}\big[\boldsymbol{\eta}_1(\mathbf{y})\big],\\
\boldsymbol{\lambda}_2^{(t)}&=\mathbb{E}_{q^{(t-1)}(\mathbf{z})}\big[\boldsymbol{\eta}_2(\mathbf{z})\big],\\
\phi_i^{(t)}&=\mathbb{E}_{q^{(t)}(\boldsymbol{\beta})}\big[\eta_i(\boldsymbol{\beta})\big],\quad i=1,\dots,n.
\end{aligned}$$

As outlined in Algorithm 2, the above expected values can be easily computed in closed form since $q(\boldsymbol{\beta})$ and $q(z_1),\dots,q(z_n)$ are already known to be Gaussian and Pólya-gammas, thereby requiring only the sequential optimization of the natural parameters. This form of CAVI, which is discussed in Hoffman et al. (2013) and is known in the literature as variational Bayesian EM (Beal and Ghahramani, 2003), clarifies the link between CAVI and the EM in Jaakkola and Jordan (2000). Indeed, recalling Section 2, both strategies optimize the same objective function and rely, implicitly, on the same steps. In particular, due to Lemma 2.1, the E-step in Algorithm 1 is in fact maximizing the conditional $\mathrm{ELBO}[q(\boldsymbol{\beta})\prod_{i=1}^n q^{(t-1)}(z_i)]$ with respect to $q(\boldsymbol{\beta})$ as in the first maximization of Algorithm 2. Similarly, the M-step solution for $\boldsymbol{\xi}$ in Algorithm 1 is actually the one maximizing the conditional $\mathrm{ELBO}[q^{(t)}(\boldsymbol{\beta})\prod_{i=1}^n q(z_i)]$ with respect to $\prod_{i=1}^n q(z_i)$ in the second optimization of the CAVI in Algorithm 2.

3.2 Stochastic Variational Inference

Algorithm 2 and model (2.4)–(2.6) motivate further generalizations in large n studies when CAVI can face severe computational bottlenecks. Indeed, each iteration of Algorithm 2 requires optimization of the whole local natural parameters ϕ_i, $i=1,\dots,n$ and a summation over the entire dataset when updating $(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)$. This issue has been addressed by Hoffman et al. (2013) via computationally cheaper updates under a SVI routine for scalable mean-field VB in conditionally conjugate exponential family models. Leveraging the probabilistic results in Section 2, we adapt this strategy to Bayesian logistic regression, thus covering an important computational gap.

To clarify the fundamental results underlying SVI, note that, by joining equations (3.1)–(3.3) and recalling Section 2.2 in Hoffman et al. (2013), the CAVI solutions for $(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)$ at iteration t are indeed those maximizing the function $\mathbb{E}_{q(\boldsymbol{\beta})}(\mathbb{E}_{q^{(t-1)}(\mathbf{z})}\{\log[p(\boldsymbol{\beta}|\mathbf{y},\mathbf{z})q(\boldsymbol{\beta})^{-1}]\})+\text{const}$, where $p(\boldsymbol{\beta}|\mathbf{y},\mathbf{z})$ and $q(\boldsymbol{\beta})$ have the same exponential family representation with natural parameters $[\boldsymbol{\eta}_1(\mathbf{y}),\boldsymbol{\eta}_2(\mathbf{z})]$ and $\boldsymbol{\lambda}=(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)$, respectively. Recalling Section 2.2 in Hoffman et al. (2013), this optimization

Algorithm 2: CAVI for logistic regression

Initialize $\xi_1^{(0)}, \ldots, \xi_n^{(0)}$.

for $t = 1$ until convergence of the evidence lower bound $\text{ELBO}[q(\boldsymbol{\beta}, \mathbf{z})]$ **do**

Maximization. Maximize $\text{ELBO}[q(\boldsymbol{\beta}) \prod_{i=1}^n q^{(t-1)}(z_i)]$ with respect to $q(\boldsymbol{\beta})$. As discussed in Section 3.1, this maximization provides a Gaussian density for $q^{(t)}(\boldsymbol{\beta})$ having natural parameters

$$\boldsymbol{\lambda}_1^{(t)} = \mathbb{E}_{q^{(t-1)}(\mathbf{z})}[\boldsymbol{\eta}_1(\mathbf{y})] = \mathbf{X}^\intercal(\mathbf{y} - 0.5 \cdot \mathbf{1}_n) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0,$$
$$\boldsymbol{\lambda}_2^{(t)} = \mathbb{E}_{q^{(t-1)}(\mathbf{z})}[\boldsymbol{\eta}_2(\mathbf{z})] = -0.5(\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^\intercal \bar{\mathbf{Z}}^{(t-1)}\mathbf{X}),$$

with $\bar{\mathbf{Z}}^{(t-1)} = \text{diag}[0.5(\xi_1^{(t-1)})^{-1}\tanh(0.5\xi_1^{(t-1)}), \ldots, 0.5(\xi_n^{(t-1)})^{-1}\tanh(0.5\xi_n^{(t-1)})]$. Hence, the approximating density is that of a $\text{N}_p(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ with $\boldsymbol{\mu}^{(t)} = (-2\boldsymbol{\lambda}_2^{(t)})^{-1}\boldsymbol{\lambda}_1^{(t)}$ and $\boldsymbol{\Sigma}^{(t)} = (-2\boldsymbol{\lambda}_2^{(t)})^{-1}$.

Maximization. Maximize $\text{ELBO}[q^{(t)}(\boldsymbol{\beta}) \prod_{i=1}^n q(z_i)]$ with respect to $\prod_{i=1}^n q(z_i)$. As discussed in Section 3.1, this maximization provides a Pólya-gamma density for each $q^{(t)}(z_i)$, $i = 1, \ldots, n$, having natural parameter

$$\phi_i^{(t)} = \mathbb{E}_{q^{(t)}(\boldsymbol{\beta})}[\eta_i(\boldsymbol{\beta})] = -0.5[\mathbf{x}_i^\intercal \boldsymbol{\Sigma}^{(t)}\mathbf{x}_i + (\mathbf{x}_i^\intercal \boldsymbol{\mu}^{(t)})^2], \quad i = 1, \ldots, n.$$

Thus, each $q^{(t)}(z_i)$ is the density of a $\text{PG}(1, \xi_i^{(t)})$ with $\xi_i^{(t)} = (-2\phi_i^{(t)})^{1/2}$. Note that $\xi_i^{(t)}$ and $-\xi_i^{(t)}$ induce the same Pólya-gamma density. Hence, there is no ambiguity in the above square root. A similar remark, from a different perspective, is found in footnote 3 of Jaakkola and Jordan (2000).

Output of the algorithm: $q^*(\boldsymbol{\beta}, \mathbf{z}) = q^*(\boldsymbol{\beta}) \prod_{i=1}^n q^*(z_i)$.

can be solved by equating to 0 the gradient

$$\nabla_{\boldsymbol{\lambda}} \mathbb{E}_{q(\boldsymbol{\beta})}\left[\log \frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})} + \sum_{i=1}^n \mathbb{E}_{q^{(t-1)}(z_i)}[\log p(y_i, z_i|\boldsymbol{\beta})]\right],$$

which coincides—under model (2.4)–(2.6)—with

$$\nabla_{\boldsymbol{\lambda}} \mathbb{E}_{q(\boldsymbol{\beta})}\left[\log \frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})}\right] + \sum_{i=1}^n \{(y_i - 0.5)\nabla_{\boldsymbol{\lambda}} \mathbb{E}_{q(\boldsymbol{\beta})}(\mathbf{x}_i^\intercal \boldsymbol{\beta})$$
$$- 0.5\mathbb{E}_{q^{(t-1)}(z_i)}(z_i)\nabla_{\boldsymbol{\lambda}} \mathbb{E}_{q(\boldsymbol{\beta})}[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2]\}.$$

This strategy leads to the estimating equations

$$\begin{aligned} &\mathbb{E}_{q^{(t-1)}(\mathbf{z})}[\boldsymbol{\eta}_1(\mathbf{y})] - \boldsymbol{\lambda}_1 = 0, \\ &\mathbb{E}_{q^{(t-1)}(\mathbf{z})}[\boldsymbol{\eta}_2(\mathbf{z})] - \boldsymbol{\lambda}_2 = 0, \end{aligned} \tag{3.4}$$

whose solution provides $\boldsymbol{\lambda}_1^{(t)}$ and $\boldsymbol{\lambda}_2^{(t)}$ as in the CAVI. See Hoffman et al. (2013) for a detailed derivation of (3.4) under a general exponential family framework.

Leveraging this view of CAVI, Hoffman et al. (2013) proposed a scalable SVI routine based on stochastic optimization (Robbins and Monro, 1951) of the ELBO in (2.3) as a direct function of the global parameters $\boldsymbol{\lambda}$. Specifically, let $q(\boldsymbol{\beta})$ the Gaussian approximating density parameterized by $\boldsymbol{\lambda}$, and $q_{\text{opt}}(z_1), \ldots, q_{\text{opt}}(z_n)$ the Pòlya-gamma densities with optimal natural parameters $\phi_1(\boldsymbol{\lambda}) = \mathbb{E}_{q(\boldsymbol{\beta})}[\eta_1(\boldsymbol{\beta})], \ldots, \phi_n(\boldsymbol{\lambda}) = \mathbb{E}_{q(\boldsymbol{\beta})}[\eta_n(\boldsymbol{\beta})]$, then optimizing the locally maximized ELBO

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\boldsymbol{\beta})}\left[\log \frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})}\right] + \sum_{i=1}^n \mathbb{E}_{q(\boldsymbol{\beta})}\{\mathbb{E}_{q_{\text{opt}}(z_i)}[\log p(y_i, z_i|\boldsymbol{\beta}) - \log q_{\text{opt}}(z_i)]\}, \tag{3.5}$$

leads to the optimal solutions for the global parameters in $\boldsymbol{\lambda}$ and, as a direct consequence, for the locally optimized ones $\phi_1(\boldsymbol{\lambda}), \ldots, \phi_n(\boldsymbol{\lambda})$. This ensures maximization of (2.3). Before deriving the SVI routine, let us highlight a key connection between the CAVI solutions in (3.4) and those arising from the optimization of $\mathcal{L}(\boldsymbol{\lambda})$. To do this, note that recalling Lemma 2.1 and its proof, the functions within the summation term in (3.5) coincide with the expected values of the conditional ELBOs in (2.1) evaluated at the optimal Pòlya-gamma densities with $\xi_i(\boldsymbol{\lambda}) = [-2\phi_i(\boldsymbol{\lambda})]^{1/2}$. In particular, $\mathbb{E}_{q(\boldsymbol{\beta})}\{\mathbb{E}_{q_{\text{opt}}(z_i)}[\log p(y_i, z_i|\boldsymbol{\beta}) - \log q_{\text{opt}}(z_i)]\}$ coincides with

$$\begin{aligned} &(y_i - 0.5)\mathbb{E}_{q(\boldsymbol{\beta})}(\mathbf{x}_i^\intercal \boldsymbol{\beta}) - 0.5\mathbb{E}_{q_{\text{opt}}(z_i)}(z_i)\{\mathbb{E}_{q(\boldsymbol{\beta})}[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2] \\ &\quad + 2\phi_i(\boldsymbol{\lambda})\} - \log\cosh\{0.5[-2\phi_i(\boldsymbol{\lambda})]^{1/2}\} + \text{const} \\ &= (y_i - 0.5)\mathbb{E}_{q(\boldsymbol{\beta})}(\mathbf{x}_i^\intercal \boldsymbol{\beta}) + \alpha[\phi_i(\boldsymbol{\lambda})] + \text{const}, \end{aligned}$$

for $i = 1, \ldots, n$, where the last equality follows by noting that $\phi_i(\boldsymbol{\lambda}) = \mathbb{E}_{q(\boldsymbol{\beta})}[\eta_i(\boldsymbol{\beta})] = \mathbb{E}_{q(\boldsymbol{\beta})}[-0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2]$

and that $-\log\cosh\{0.5[-2\phi_i(\boldsymbol{\lambda})]^{1/2}\}$ defines the function $\alpha[\phi_i(\boldsymbol{\lambda})]$ in the exponential family representation for the density of the Pòlya-gamma with parameters 1 and $[-2\phi_i(\boldsymbol{\lambda})]^{1/2}$. Since our final goal is to maximize $\mathcal{L}(\boldsymbol{\lambda})$, let us substitute the above equation in (3.5) and compute $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$. This leads to

$$\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{q(\boldsymbol{\beta})}\left[\log\frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})}\right]+\sum_{i=1}^{n}\{(y_i-0.5)\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{q(\boldsymbol{\beta})}(\mathbf{x}_i^\intercal\boldsymbol{\beta}) + \nabla_{\phi_i(\boldsymbol{\lambda})}\alpha[\phi_i(\boldsymbol{\lambda})]\nabla_{\boldsymbol{\lambda}}\phi_i(\boldsymbol{\lambda})\}=\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}),$$

where

$$\nabla_{\phi_i(\boldsymbol{\lambda})}\alpha[\phi_i(\boldsymbol{\lambda})]\nabla_{\boldsymbol{\lambda}}\phi_i(\boldsymbol{\lambda}) = -0.5\mathbb{E}_{q_{\text{opt}}(z_i)}(z_i)\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{q(\boldsymbol{\beta})}[(\mathbf{x}_i^\intercal\boldsymbol{\beta})^2].$$

To clarify the above expression, recall that $\phi_i(\boldsymbol{\lambda}) = \mathbb{E}_{q(\boldsymbol{\beta})}[-0.5(\mathbf{x}_i^\intercal\boldsymbol{\beta})^2]$, whereas from the properties of exponential families we also have that $\nabla_{\phi_i(\boldsymbol{\lambda})}\alpha[\phi_i(\boldsymbol{\lambda})] = \mathbb{E}_{q_{\text{opt}}(z_i)}(z_i)$. This expression for $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$ provides an optimization partially related to CAVI. Indeed, by comparing the above gradient with the one leading to equations (3.4), it can be noticed that these quantities coincide after replacing $\mathbb{E}_{q^{(t-1)}(z_i)}(z_i)$ with $\mathbb{E}_{q_{\text{opt}}(z_i)}(z_i)$. Hence, the maximum of $\mathcal{L}(\boldsymbol{\lambda})$ can be obtained solving equations (3.4), where the expected value is now computed with respect to $q_{\text{opt}}(\mathbf{z})$ instead of $q^{(t-1)}(\mathbf{z})$.

To derive the SVI algorithm, let us first re-express $\mathbb{E}_{q_{\text{opt}}(\mathbf{z})}[\boldsymbol{\eta}_1(\mathbf{y})]-\boldsymbol{\lambda}_1=0$ and $\mathbb{E}_{q_{\text{opt}}(\mathbf{z})}[\boldsymbol{\eta}_2(\mathbf{z})]-\boldsymbol{\lambda}_2=0$ as

$$\begin{aligned}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\sum_{i=1}^{n}\mathbf{x}_i(y_i-0.5)-\boldsymbol{\lambda}_1=0,\\ -0.5\left[\boldsymbol{\Sigma}_0^{-1}+\sum_{i=1}^{n}\mathbf{x}_i\mathbb{E}_{q_{\text{opt}}(z_i)}(z_i)\mathbf{x}_i^\intercal\right]-\boldsymbol{\lambda}_2=0,\end{aligned}\tag{3.6}$$

to highlight how the evaluation of (3.6) requires storing the entire dataset and summing over all the units. Such a step could be a major computational bottleneck when the sample size n is massive, thereby motivating the optimization of $\mathcal{L}(\boldsymbol{\lambda})$ (Hoffman et al., 2013) via stochastic approximation of (3.6) (Robbins and Monro, 1951). This is obtained by constructing a random version of $\{\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+\sum_{i=1}^{n}\mathbf{x}_i(y_i-0.5)-\boldsymbol{\lambda}_1, -0.5[\boldsymbol{\Sigma}_0^{-1}+\sum_{i=1}^{n}\mathbf{x}_i\mathbb{E}_{q_{\text{opt}}(z_i)}(z_i)\mathbf{x}_i^\intercal]-\boldsymbol{\lambda}_2\}$ whose expected value coincides with these functions, but its realizations are cheaper to compute. A simple solution is to rely on the discrete random variable $B(\boldsymbol{\lambda})$ taking values $\{B_i(\boldsymbol{\lambda}_1) = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + n\mathbf{x}_i(y_i-0.5)-\boldsymbol{\lambda}_1, B_i(\boldsymbol{\lambda}_2) = -0.5[\boldsymbol{\Sigma}_0^{-1}+n\mathbf{x}_i\mathbb{E}_{q_{\text{opt}}(z_i)}(z_i)\mathbf{x}_i^\intercal]-\boldsymbol{\lambda}_2\}$, for $i=1,\dots,n$ with equal probability n^{-1}, thus implicitly relying on a mechanism which samples a unit i uniformly and then computes (3.6) as if such a unit was observed n times. This allows the application of Robbins and Monro (1951) to solve (3.6) via the iterative updates

$$\begin{aligned}\boldsymbol{\lambda}_1^{(t)}&=\boldsymbol{\lambda}_1^{(t-1)}+\rho_t B_t(\boldsymbol{\lambda}_1^{(t-1)}),\\ \boldsymbol{\lambda}_2^{(t)}&=\boldsymbol{\lambda}_2^{(t-1)}+\rho_t B_t(\boldsymbol{\lambda}_2^{(t-1)}),\end{aligned}\tag{3.7}$$

for every iteration t, where $[B_t(\boldsymbol{\lambda}_1^{(t-1)}), B_t(\boldsymbol{\lambda}_2^{(t-1)})]$ denotes an independent draw from $B(\boldsymbol{\lambda})$, evaluated at $(\boldsymbol{\lambda}_1^{(t-1)}, \boldsymbol{\lambda}_2^{(t-1)})$, whereas ρ_t characterize step-sizes ensuring convergence to the solution of (3.6)—and hence to the maximum of $\mathcal{L}(\boldsymbol{\lambda})$—when $\sum_t\rho_t=+\infty$ and $\sum_t\rho_t^2<+\infty$ (Robbins and Monro, 1951, Spall, 2003). Hoffman et al. (2013) set $\rho_t=(t+\tau)^{-\kappa}$, with $\kappa\in(0.5,1]$ denoting the forgetting rate, and $\tau\geq 0$ the delay down-weighting early iterations. These settings ensure the convergence conditions on ρ_t. Algorithm 3 provides the pseudo-code to perform SVI in logistic regression under model (2.4)–(2.6). As it can be noticed, this routine relies on updating steps which are cheaper to compute than those of CAVI. In fact, each iteration of Algorithm 3 does not require to sum over the entire dataset, but relies instead on a single observation sampled uniformly. These gains are fundamental to scale-up calculations in massive datasets.

Figure 1 provides a summarizing quantitative assessment for the performance of CAVI and SVI in the logistic regression with $\text{logit}[\text{pr}(y_i=1|\boldsymbol{\beta})]=\beta_1+\beta_2 x_i$ for each $i=1,\dots,n$. To study the performance under different dimensions, we generate data for an increasing sample size $n\in(20,100,1000,10{,}000)$ from a logistic regression with true coefficients set equal to 1 and covariates $x_1,\dots,x_n$ from a UNIF$(-2,2)$. We perform Bayesian inference under a moderately diffuse prior $\boldsymbol{\beta}\sim\text{N}_2(\mathbf{0},10\cdot\mathbf{I}_2)$ and approximate the posterior via CAVI and SVI, with $(\tau,\kappa)=(1,0.75)$. As highlighted in Figure 1, although SVI is based on noisy gradients, the final approximations $q^*_{\text{SVI}}(\beta_1)$ and $q^*_{\text{SVI}}(\beta_2)$ are similar to the optimal solutions $q^*(\beta_1)$ and $q^*(\beta_2)$ from the CAVI. These approximate posteriors increasingly shrink around the true coefficients when n grows, thus suggesting desirable asymptotic behavior of the CAVI and SVI solutions. Code and tutorials to reproduce this analysis are available at https://github.com/tommasorigon/logisticVB.

4. DISCUSSION

Motivated by the success of the lower bound developed by Jaakkola and Jordan (2000) for logistic log-likelihoods, and by the lack of formal justifications for its excellent performance, we introduced a novel connection between their construction and a Pólya-gamma

Algorithm 3: SVI for logistic regression

Initialize $(\boldsymbol{\lambda}_1^{(0)}, \boldsymbol{\lambda}_2^{(0)})$ randomly and set the step-size sequence ρ_t appropriately.

for $t = 1$ until a large number of iterations (in the illustrative application we rely on 10^4 iterations) **do**

Sampling. Sample a data point $(y_i, \mathbf{x}_i)$ randomly from the dataset.

Local maximization. Calculate the locally optimized density for z_i as a function of the latest value $\boldsymbol{\lambda}^{(t-1)}$ for $\boldsymbol{\lambda}$, thus obtaining a Pólya-gamma with natural parameter

$$\phi_i(\boldsymbol{\lambda}^{(t-1)}) = -0.5[\mathbf{x}_i^\intercal \boldsymbol{\Sigma}^{(t-1)} \mathbf{x}_i + (\mathbf{x}_i^\intercal \boldsymbol{\mu}^{(t-1)})^2] = -0.5\{\mathbf{x}_i^\intercal (-2\boldsymbol{\lambda}_2^{(t-1)})^{-1} \mathbf{x}_i + [\mathbf{x}_i^\intercal (-2\boldsymbol{\lambda}_2^{(t-1)})^{-1} \boldsymbol{\lambda}_1^{(t-1)}]^2\}.$$

Therefore, the optimal solution is the density of a $\text{PG}[1, \xi_i(\boldsymbol{\lambda}^{(t-1)})]$, with $\xi_i(\boldsymbol{\lambda}^{(t-1)}) = [-2\phi_i(\boldsymbol{\lambda}^{(t-1)})]^{1/2}$.

Global parameters updates. Update the global parameters according to Robbins and Monro (1951) iterative procedure outlined in (3.7). This approach provides the solutions

$$\boldsymbol{\lambda}_1^{(t)} = (1-\rho_t)\boldsymbol{\lambda}_1^{(t-1)} + \rho_t[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + n\mathbf{x}_i(y_i - 0.5)],$$

$$\boldsymbol{\lambda}_2^{(t)} = (1-\rho_t)\boldsymbol{\lambda}_2^{(t-1)} - \rho_t 0.5(\boldsymbol{\Sigma}_0^{-1} + n\mathbf{x}_i \bar{z}_i^{(t-1)} \mathbf{x}_i^\intercal),$$

with $\bar{z}_i^{(t-1)} = 0.5[\xi_i(\boldsymbol{\lambda}^{(t-1)})]^{-1} \tanh[0.5\xi_i(\boldsymbol{\lambda}^{(t-1)})]$. Hence, the approximating density is that of a Gaussian with mean $\boldsymbol{\mu}^{(t)} = (-2\boldsymbol{\lambda}_2^{(t)})^{-1}\boldsymbol{\lambda}_1^{(t)}$ and variance-covariance matrix $\boldsymbol{\Sigma}^{(t)} = (-2\boldsymbol{\lambda}_2^{(t)})^{-1}$.

Output of the algorithm: $q^*_{\text{SVI}}(\boldsymbol{\beta})$.

data augmentation developed in the recent years for logistic regression (Polson, Scott and Windle, 2013). Besides providing a probabilistic interpretation of the bound derived by Jaakkola and Jordan (2000), this connection crucially places the variational methods associated with the proposed lower bound within a more general framework having desirable properties. More specifically, the EM for variational inference proposed by Jaakkola and Jordan (2000) maximizes a genuine ELBO associated with a conditionally conjugate exponential family model and, hence, provides the same approximation of the CAVI for VB in this model.

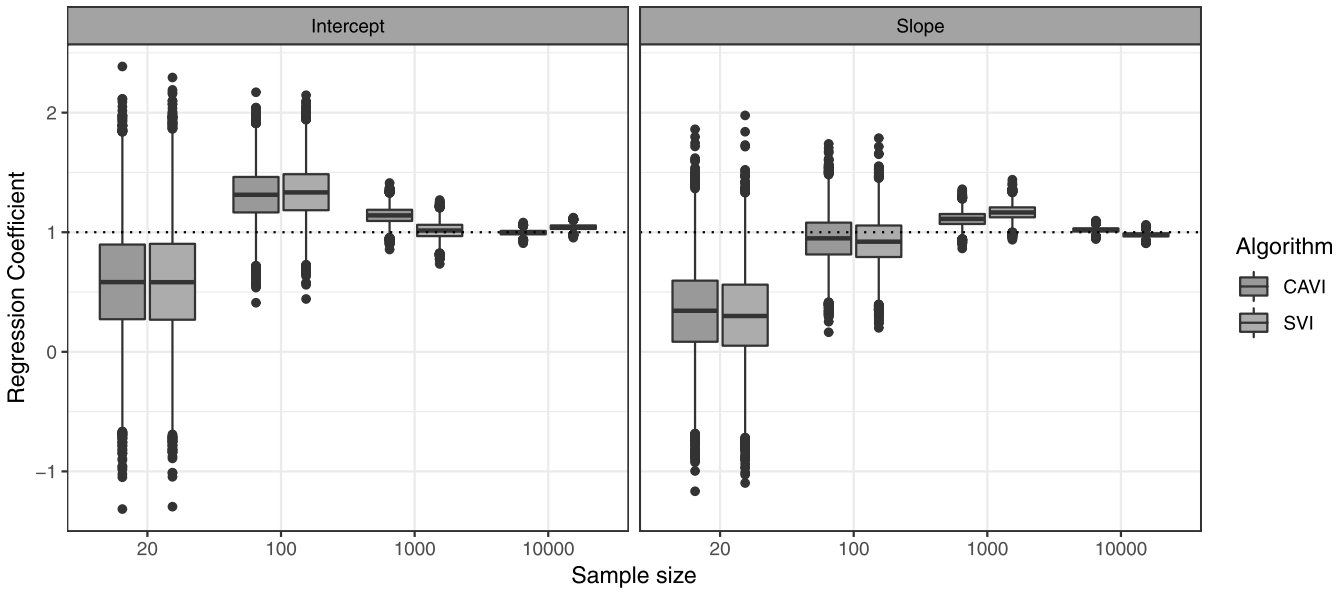

FIG. 1. *For increasing n, boxplots of the* CAVI *and* SVI *solutions for the coefficients in a Bayesian logistic regression with a single covariate. The dotted horizontal line refers to the true coefficients. The boxplots are constructed using 10^4 samples from the optimal Gaussian approximations under the different variational strategies and sample sizes.*

The above finding motivates further generalizations to novel computational methods, including the SVI algorithm in Section 3.2. On a similar line of research, an interesting direction is to incorporate the methods of Giordano, Broderick and Jordan (2015) to correct the variance-covariance matrix in $q^*(\boldsymbol{\beta})$ from Algorithms 1–3, which is known to underestimate variability. Besides this, the results in Figure 1 motivate also future theoretical studies on the quality of the CAVI and SVI approximations in asymptotic settings. This can be done by adapting the available theory on mean-field VB for conditionally conjugate exponential family models (e.g., Wang and Titterington, 2004). Finally, we shall also emphasize that although our focus is on classical Bayesian logistic regression, the results in Sections 2–3 can be easily generalized to more complex learning procedures incorporating logistic models as a building block, as long as such formulations admit conditionally conjugate exponential family representations.

APPENDIX: MAXIMUM LIKELIHOOD ESTIMATION

Although maximum likelihood estimation for the coefficients of a logistic regression is well established, there is still active research within this class of models to address other important open questions. For instance, the classical Newton–Raphson does not guarantee monotone log-likelihood sequences, thus potentially affecting the stability of the maximization routine (Böhning and Lindsay, 1988). This issue has motivated other methods leveraging alternative quadratic approximations which uniformly minorize the logistic log-likelihood while being tangent to it (Böhning and Lindsay, 1988; de Leeuw and Lange, 2009; Browne and McNicholas, 2015), thus guaranteeing monotone convergence (Hunter and Lange, 2004). As discussed in Sections 1–2, this is the case of the bound (1.6) in Jaakkola and Jordan (2000).

Motivated by this result, Jaakkola and Jordan (2000) provided in Appendix C of their article an iterative routine for maximum likelihood estimation of $\boldsymbol{\beta}$ that has monotone log-likelihood sequences and simple maximizations. In particular, letting $\boldsymbol{\beta}^{(t-1)}$ be the estimate of the coefficients at the iteration $t-1$ and simplifying the calculations in Appendix C of Jaakkola and Jordan (2000), their routine first maximizes (1.6) with respect to each $\xi_1, \dots, \xi_n$, obtaining $\hat{\xi}_i^{(t-1)} = \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t-1)}$, for $i = 1, \dots, n$, and then derive $\boldsymbol{\beta}^{(t)}$ by maximizing

$$\log \bar{p}^{(t-1)}(\mathbf{y}|\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \bar{p}^{(t-1)}(y_i|\boldsymbol{\beta}),$$

with ξ_i replaced by $\hat{\xi}_i^{(t-1)} = \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t-1)}$. This last optimization is straightforward due to the quadratic form of (1.6), thus providing

$$\boldsymbol{\beta}^{(t)} = (\mathbf{X}^\intercal \hat{\mathbf{Z}}^{(t-1)} \mathbf{X})^{-1} \mathbf{X}^\intercal (\mathbf{y} - 0.5 \cdot \mathbf{1}_n), \tag{A.1}$$

where $\hat{\mathbf{Z}}^{(t-1)}$ denotes a diagonal matrix having entries $\hat{\mathbf{Z}}^{(t-1)}_{[ii]} = 0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t-1)})^{-1} \tanh[0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t-1)})]$.

Analyzing such a strategy in the light of (2.2), it can be noticed that $\hat{\xi}_i^{(t-1)}$ leads to the solution $\hat{q}^{(t-1)}(z_i)$ minimizing the KL divergence $\mathrm{KL}[q(z_i) \| p(z_i|\boldsymbol{\beta}^{(t-1)})]$ in (2.2), for $i = 1, \dots, n$, whereas the function

$$\begin{aligned} \log \bar{p}^{(t-1)}(\mathbf{y}|\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log \bar{p}^{(t-1)}(y_i|\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \mathbb{E}_{\hat{q}^{(t-1)}(z_i)} \left[\log \frac{p(y_i, z_i|\boldsymbol{\beta})}{\hat{q}^{(t-1)}(z_i)} \right], \end{aligned}$$

maximized with respect to the parameters comprising $\boldsymbol{\beta}$ is equal, up to an additive constant, to the expectation $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t-1)})$ of the complete log-likelihood function $\log p(\mathbf{y}, \mathbf{z}|\boldsymbol{\beta}) = \sum_{i=1}^{n} \log p(y_i, z_i|\boldsymbol{\beta})$ computed with respect to the conditional distribution of the augmented Pólya-gamma data $(z_i|\boldsymbol{\beta}^{(t-1)}) \sim \mathrm{PG}(1, \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t-1)})$, for $i = 1, \dots, n$. Combining these results with the EM rationale (Bishop, 2006, Chapter 9.4), it follows that the routine in Appendix C of Jaakkola and Jordan (2000) is an EM based on Pólya-gamma augmented data. This algorithm first computes the expected value $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t-1)}) = \sum_{i=1}^{n} \mathbb{E}_{\hat{q}^{(t-1)}(z_i)}[\log p(y_i, z_i|\boldsymbol{\beta})]$ and then maximizes it with respect to $\boldsymbol{\beta}$. See also Scott and Sun (2013).

As discussed by Jaakkola and Jordan (2000), the above maximization strategy guarantees a monotone log-likelihood sequence, ensuring stable convergence. Indeed, leveraging equations (1.6) and (2.2), it can be noticed that this routine provides a minorize-maximize (MM) algorithm (e.g., Hunter and Lange, 2004), provided that $\sum_{i=1}^{n} \log p(y_i|\boldsymbol{\beta}) \geq \sum_{i=1}^{n} \log \bar{p}^{(t-1)}(y_i|\boldsymbol{\beta})$ for every vector $\boldsymbol{\beta}$, and that $\sum_{i=1}^{n} \log p(y_i|\boldsymbol{\beta}^{(t-1)}) = \sum_{i=1}^{n} \log \bar{p}^{(t-1)}(y_i|\boldsymbol{\beta}^{(t-1)})$. We shall notice that also de Leeuw and Lange (2009) and Browne and McNicholas (2015) highlighted this relation with the MM under a mathematical argument and discussed the sharpness of (1.6). Exploiting results in Section 2, we also show that the MM algorithm relying on (1.6) improves the convergence rate of the one in Böhning and Lindsay (1988). To our knowledge, this is the only tractable MM alternative to Jaakkola and Jordan (2000).

In addressing the above goal, let us first rewrite (A.1) to facilitate a direct comparison with the solution from Böhning and Lindsay (1988), given by

$$\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)} + (\mathbf{X}^\intercal \boldsymbol{\Gamma} \mathbf{X})^{-1} \mathbf{X}^\intercal (\mathbf{y} - \boldsymbol{\pi}^{(t-1)}), \tag{A.2}$$

where $\boldsymbol{\Gamma} = 0.25 \cdot \mathbf{I}_n$ and $\boldsymbol{\pi}^{(t-1)} = (\pi_1^{(t-1)}, \ldots, \pi_n^{(t-1)})^\intercal$ with

$$\pi_i^{(t-1)} = [1 + \exp(-\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t-1)})]^{-1}.$$

Indeed, adding and subtracting $\hat{\mathbf{Z}}^{(t-1)} \mathbf{X} \boldsymbol{\beta}^{(t-1)}$ in $(\mathbf{y} - 0.5 \cdot \mathbf{1}_n)$, equation (A.1) reduces to

$$\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)} + (\mathbf{X}^\intercal \hat{\mathbf{Z}}^{(t-1)} \mathbf{X})^{-1} \mathbf{X}^\intercal (\mathbf{y} - \boldsymbol{\pi}^{(t-1)}), \tag{A.3}$$

after noticing that every single element $y_i - 0.5 - 0.5 \tanh[0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t-1)})]$ in $\mathbf{y} - 0.5 \cdot \mathbf{1}_n - \hat{\mathbf{Z}}^{(t-1)} \mathbf{X} \boldsymbol{\beta}^{(t-1)}$ can be alternatively re-expressed as

$$y_i - 0.5\left[1 + \frac{1 - \exp(-\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t-1)})}{1 + \exp(-\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t-1)})}\right] = y_i - \pi_i^{(t-1)}.$$

A closer inspection of the two equations in (A.2) and (A.3) shows that the updating underlying Böhning and Lindsay (1988) and Jaakkola and Jordan (2000) coincides with the one arising from the Newton–Raphson, after replacing the Hessian $\mathbf{H}^{(t-1)} = -\mathbf{X}^\intercal \boldsymbol{\Lambda}^{(t-1)} \mathbf{X} = -\mathbf{X}^\intercal \mathrm{diag}[\pi_1^{(t-1)}(1-\pi_1^{(t-1)}), \ldots, \pi_n^{(t-1)}(1-\pi_n^{(t-1)})] \mathbf{X}$ of the logistic log-likelihood, with $-\mathbf{X}^\intercal \boldsymbol{\Gamma} \mathbf{X}$ in (A.2) and $-\mathbf{X}^\intercal \hat{\mathbf{Z}}^{(t-1)} \mathbf{X}$ in (A.3). Recalling Böhning and Lindsay (1988), both matrices define a lower bound for the Hessian and guarantee that the updates in (A.2) and (A.3) induce a monotone sequence for the log-likelihood. In Böhning and Lindsay (1988) the uniform bound follows after noticing that $\boldsymbol{\pi}^{(t-1)}(1 - \boldsymbol{\pi}^{(t-1)}) \leq 0.25 \cdot \mathbf{1}_n$ for any $\boldsymbol{\pi}^{(t-1)} \in (0, 1)^n$, whereas, according to Lemma 2.1, the adaptive bound induced by Jaakkola and Jordan (2000) is formally related to an exact data augmentation, thus suggesting that (A.3) may provide more efficient updates than (A.2). This claim is formalized in Proposition A.1 by comparing the convergence rates of the two algorithms. Refer to McLachlan and Krishnan (1997), Chapter 3.9, for details regarding the definition and the computation of the convergence rate associated with a generic iterative routine.

PROPOSITION A.1. *Assume that $\boldsymbol{\beta}^*$ characterizes the limit, if it exists, of $\{\boldsymbol{\beta}^{(t)} : t \geq 1\}$, and let $M^{\mathrm{B}}(\cdot) = \{M_1^{\mathrm{B}}(\cdot), \ldots, M_p^{\mathrm{B}}(\cdot)\}$ and $M^{\mathrm{J}}(\cdot) = \{M_1^{\mathrm{J}}(\cdot), \ldots, M_p^{\mathrm{J}}(\cdot)\}$ denote the functions which update the estimate of $\boldsymbol{\beta}$ at iteration $t-1$ to that at iteration t in (A.2) and (A.3), respectively. Then $r_{\mathrm{B}} \geq r_{\mathrm{J}}$, with $r_{\mathrm{B}} = \|\boldsymbol{J}_{\mathrm{B}}^*\|_2$ and $r_{\mathrm{J}} = \|\boldsymbol{J}_{\mathrm{J}}^*\|_2$ being the maximum eigenvalues of the Jacobians $\boldsymbol{J}_{\mathrm{B}} = \partial M^{\mathrm{B}}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ and $\boldsymbol{J}_{\mathrm{J}} = \partial M^{\mathrm{J}}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, respectively, computed in $\boldsymbol{\beta}^*$.*

To prove Proposition A.1, first note that $\boldsymbol{J}_{\mathrm{B}}^* = \mathbf{I}_p + (\mathbf{X}^\intercal \boldsymbol{\Gamma} \mathbf{X})^{-1} \mathbf{H}^*$ can be easily computed as in Böhning and Lindsay (1988), since $\boldsymbol{\Gamma}$ does not depend on $\boldsymbol{\beta}$ in (A.2). It is instead not immediate to calculate $\boldsymbol{J}_{\mathrm{J}}^*$ via direct differentiation of $\mathrm{M}^{\mathrm{J}}(\boldsymbol{\beta})$, because (A.3) contains more complex hyperbolic transformations of $\boldsymbol{\beta}$. However, exploiting the probabilistic findings in Section 2, this issue can be easily circumvented by leveraging the EM interpretation of the routine in Jaakkola and Jordan (2000) via Pólya-gamma augmented data. Indeed, following McLachlan and Krishnan (1997), Chapter 3.9.3, the rate matrix of an iterative routine relying on EM methods, coincides with $\boldsymbol{J}_{\mathrm{J}}^* = \mathbf{I}_p + \mathcal{I}_c(\boldsymbol{\beta}^*)^{-1} \mathbf{H}^*$, where $\mathcal{I}_c(\boldsymbol{\beta}^*)$ denotes the expectation, taken with respect to the augmented data, of the complete-data information matrix $\mathrm{I}_c(\boldsymbol{\beta}^*)$. This quantity can be easily computed in our case, provided that the complete log-likelihood is equal, up to an additive constant, to the quadratic function $\sum_{i=1}^n [(y_i - 0.5)\mathbf{x}_i^\intercal \boldsymbol{\beta} - 0.5 z_i (\mathbf{x}_i^\intercal \boldsymbol{\beta})^2]$ of $\boldsymbol{\beta}$, which is also linear in the augmented Pólya-gamma data z_i. Due to this, it is easy to show that $\mathcal{I}_c(\boldsymbol{\beta}^*) = \mathbf{X}^\intercal \hat{\mathbf{Z}}^* \mathbf{X}$, where $\hat{\mathbf{Z}}^*$ denotes the $n \times n$ diagonal matrix with entries $\hat{\mathbf{Z}}_{[ii]}^* = 0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta}^*)^{-1} \tanh[0.5(\mathbf{x}_i^\intercal \boldsymbol{\beta}^*)]$.

PROOF. Recalling the above discussion, the proof of Proposition A.1 requires comparing the maximum eigenvalues of

$$\begin{aligned} \boldsymbol{J}_{\mathrm{B}}^* &= \mathbf{I}_p + (\mathbf{X}^\intercal \boldsymbol{\Gamma} \mathbf{X})^{-1} \mathbf{H}^* \\ &= \mathbf{I}_p - (\mathbf{X}^\intercal \boldsymbol{\Gamma} \mathbf{X})^{-1} (\mathbf{X}^\intercal \boldsymbol{\Lambda}^* \mathbf{X}) \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{J}_{\mathrm{J}}^* &= \mathbf{I}_p + (\mathbf{X}^\intercal \hat{\mathbf{Z}}^* \mathbf{X})^{-1} \mathbf{H}^* \\ &= \mathbf{I}_p - (\mathbf{X}^\intercal \hat{\mathbf{Z}}^* \mathbf{X})^{-1} (\mathbf{X}^\intercal \boldsymbol{\Lambda}^* \mathbf{X}). \end{aligned}$$

To address this goal, first note that

$$\begin{aligned} \|\boldsymbol{J}_{\mathrm{B}}^*\|_2 &= 1 - \|(\mathbf{X}^\intercal \boldsymbol{\Gamma} \mathbf{X})^{-1} (\mathbf{X}^\intercal \boldsymbol{\Lambda}^* \mathbf{X})\|_2, \\ \|\boldsymbol{J}_{\mathrm{J}}^*\|_2 &= 1 - \|(\mathbf{X}^\intercal \hat{\mathbf{Z}}^* \mathbf{X})^{-1} (\mathbf{X}^\intercal \boldsymbol{\Lambda}^* \mathbf{X})\|_2. \end{aligned}$$

Therefore, $r_{\mathrm{B}} \geq r_{\mathrm{J}}$ if $\|(\mathbf{X}^\intercal \hat{\mathbf{Z}}^* \mathbf{X})^{-1} (\mathbf{X}^\intercal \boldsymbol{\Lambda}^* \mathbf{X})\|_2 \geq \|(\mathbf{X}^\intercal \boldsymbol{\Gamma} \mathbf{X})^{-1} (\mathbf{X}^\intercal \boldsymbol{\Lambda}^* \mathbf{X})\|_2$. This inequality can be proved

by noticing that

$$\begin{aligned}&\|(\mathbf{X}^\intercal\boldsymbol{\Gamma}\mathbf{X})^{-1}(\mathbf{X}^\intercal\boldsymbol{\Lambda}^*\mathbf{X})\|_2\\&\quad=\|(\mathbf{X}^\intercal\boldsymbol{\Gamma}\mathbf{X})^{-1}(\mathbf{X}^\intercal\hat{\mathbf{Z}}^*\mathbf{X})(\mathbf{X}^\intercal\hat{\mathbf{Z}}^*\mathbf{X})^{-1}(\mathbf{X}^\intercal\boldsymbol{\Lambda}^*\mathbf{X})\|_2\\&\quad\le\|(\mathbf{X}^\intercal\boldsymbol{\Gamma}\mathbf{X})^{-1}(\mathbf{X}^\intercal\hat{\mathbf{Z}}^*\mathbf{X})\|_2\|(\mathbf{X}^\intercal\hat{\mathbf{Z}}^*\mathbf{X})^{-1}(\mathbf{X}^\intercal\boldsymbol{\Lambda}^*\mathbf{X})\|_2.\end{aligned}$$

Hence, a sufficient condition for the above inequality to be true is that $\|(\mathbf{X}^\intercal\boldsymbol{\Gamma}\mathbf{X})^{-1}(\mathbf{X}^\intercal\hat{\mathbf{Z}}^*\mathbf{X})\|_2 \le 1$ or, alternatively, that $\hat{\mathbf{Z}}^*_{[ii]} = 0.5(\mathbf{x}_i^\intercal\boldsymbol{\beta}^*)^{-1}\tanh[0.5(\mathbf{x}_i^\intercal\boldsymbol{\beta}^*)] \le 0.25$ for each $i = 1,\dots,n$; recall also that $\boldsymbol{\Gamma} = 0.25\cdot\mathbf{I}_n$. Letting $u = 0.5(\mathbf{x}_i^\intercal\boldsymbol{\beta}^*)$, and rewriting the inequality $u^{-1}\sinh(u) \le \cosh(u)$ (Zhu, 2012) as $u^{-1}\tanh(u) \le 1$, it directly follows that $2(\mathbf{x}_i^\intercal\boldsymbol{\beta}^*)^{-1}\tanh[0.5(\mathbf{x}_i^\intercal\boldsymbol{\beta}^*)] \le 1$, thus guaranteeing that $0.5(\mathbf{x}_i^\intercal\boldsymbol{\beta}^*)^{-1}\tanh[0.5(\mathbf{x}_i^\intercal\boldsymbol{\beta}^*)] \le 0.25$. This concludes the proof. □

Proposition A.1 ensures that the mapping in (A.3) improves the convergence rate of (A.2). In fact, higher values of r imply slower convergence. We shall however emphasize that the EM in Appendix C of Jaakkola and Jordan (2000) does not reach the quadratic convergence of the Newton–Raphson strategy, but guarantees a monotone log-likelihood sequence. It is also important to highlight that although the MM in Böhning and Lindsay (1988) has slower convergence, the matrix $\mathbf{X}^\intercal\boldsymbol{\Gamma}\mathbf{X}$ in (A.2) does not depend on $\boldsymbol{\beta}^{(t-1)}$, thus requiring inversion only once during the iterative procedure. This result reduces computational complexity, especially in high-dimensional problems, compared to the updating in (A.3), which requires, instead, the inversion of $\mathbf{X}^\intercal\hat{\mathbf{Z}}^{(t-1)}\mathbf{X}$ at every iteration. We refer to the tutorial `em_logistic_tutorial.md` in https://github.com/tommasorigon/logisticVB for illustrative simulations.

Although the above focus has been on the maximum likelihood estimation method, the probabilistic interpretation (2.1) of the quadratic bound in Jaakkola and Jordan (2000) motivates simple adaptations to include the maximum a posteriori estimation problem under a Bayesian framework. This routine has been carefully studied by Scott and Sun (2013) and we refer to their contribution for details.

ACKNOWLEDGEMENTS

The authors are grateful to the Editor, the Associate Editor and the referees for the valuable comments and suggestions, which helped in substantially improving the initial version of this article.

REFERENCES

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

BEAL, M. J. and GHAHRAMANI, Z. (2003). The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 453–463. Oxford Univ. Press, New York. MR2003189

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer, New York. MR2247587

BISHOP, C. M. and SVENSÉN, M. (2003). Bayesian hierarchical mixtures of experts. *Proc. Conf. Uncertain. Artif. Intell.* 57–64.

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

BÖHNING, D. and LINDSAY, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Ann. Inst. Statist. Math.* **40** 641–663. MR0996690

BRAUN, M. and MCAULIFFE, J. (2010). Variational inference for large-scale models of discrete choice. *J. Amer. Statist. Assoc.* **105** 324–335. MR2757203

BROWNE, R. P. and MCNICHOLAS, P. D. (2015). Multivariate sharp quadratic bounds via Σ-strong convexity and the Fenchel connection. *Electron. J. Stat.* **9** 1913–1938. MR3391124

CARBONETTO, P. and STEPHENS, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **7** 73–107. MR2896713

CHOI, H. M. and HOBERT, J. P. (2013). The Polya-gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electron. J. Stat.* **7** 2054–2064. MR3091616

DE LEEUW, J. and LANGE, K. (2009). Sharp quadratic majorization in one dimension. *Comput. Statist. Data Anal.* **53** 2471–2484. MR2665900

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

GIORDANO, R. J., BRODERICK, T. and JORDAN, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. *Adv. Neural Inf. Process. Syst.* 1441–1449.

HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14** 1303–1347. MR3081926

HUNTER, D. R. and LANGE, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509

JAAKKOLA, T. S. and JORDAN, M. I. (2000). Bayesian parameter estimation via variational methods. *Stat. Comput.* **10** 25–37.

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.

KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22** 79–86. MR0039968

LEE, S., HUANG, J. Z. and HU, J. (2010). Sparse logistic principal components analysis for binary data. *Ann. Appl. Stat.* **4** 1579–1601. MR2758342

MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1417721

ORMEROD, J. T. and WAND, M. P. (2010). Explaining variational approximations. *Amer. Statist.* **64** 140–153. MR2757005

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

REN, L., DU, L., CARIN, L. and DUNSON, D. B. (2011). Logistic stick-breaking process. *J. Mach. Learn. Res.* **12** 203–239. MR2773552

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668

SCOTT, J. G. and SUN, L. (2013). Expectation-maximization for logistic regression. Available at arXiv:1306.0040.

SPALL, J. C. (2003). *Introduction to Stochastic Search and Optimization*: *Estimation*, *Simulation*, *and Control*. *Wiley-Interscience Series in Discrete Mathematics and Optimization*. Wiley Interscience, Hoboken, NJ. MR1968388

TANG, Y., BROWNE, R. P. and MCNICHOLAS, P. D. (2015). Model based clustering of high-dimensional binary data. *Comput. Statist. Data Anal.* **87** 84–101. MR3319809

WAND, M. P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *J. Amer. Statist. Assoc.* **112** 137–156. MR3646558

WAND, M. P., ORMEROD, J. T., PADOAN, S. A. and FRÜHRWIRTH, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Anal.* **6** 847–900. MR2869967

WANG, C. and BLEI, D. M. (2013). Variational inference in nonconjugate models. *J. Mach. Learn. Res.* **14** 1005–1031. MR3063617

WANG, B. and TITTERINGTON, D. M. (2004). Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. *Proc. Conf. Uncertain. Artif. Intell.* 577–584.

ZHU, L. (2012). New inequalities for hyperbolic functions and their applications. *J. Inequal. Appl.* **303** 1–29. MR3017334