UNIVERSITA' COMMERCIALE "LUIGI BOCCONI"
PhD SCHOOL

PhD program in Statistics
Cycle: XXXIV
Disciplinary Field: SECS-S/01

Methodological and Computational Advances for High–Dimensional Bayesian Regression with Binary and Categorical Responses

Advisor: Daniele DURANTE
Co-Advisor: Giacomo ZANELLA

PhD Thesis by
Niccolò ANCESCHI
ID number: 3084435

Year 2023

To Anas

Acknowledgements

At the end of such an important step in my professional and personal growth, I cannot but express my sincere gratitude to all who supported me along this track.
First of all, my Advisors Daniele and Giacomo. Their dedication and competence have always motivated and guided me in my research work, together with their constant encouragement and the confidence they conveyed to me. My thanks go also to the whole Bocconi Faculty, for their teaching and mentorship in introducing me to the fascinating research field of Bayesian statistics.
I am also grateful for having shared this journey with all my colleagues Ph.D. students, in particular Francesco, Giovanni, Marta, Augusto, Laura, Beatrice, Stefano, and Tommaso. From the passion and enthusiasm for hour work to the daily life together, I really believe we shared something special and out of ordinary. I cannot but hope that this path will continue together.
A special thought is reserved for all those who supported me in recovering from the motorbike accident that marked so profoundly my Ph.D. track. From the doctors to the paramedics, from the nursing staff to the physiotherapists, you literally got me back on my feet.
Finally, I want to deeply thank my family, my friends, and my girlfriend Giudi for being always by my side on this path as in all the rest.

Abstract

Probit and logistic regressions are among the most popular and well–established formulations to model binary observations, thanks to their plain structure and high interpretability. Despite their simplicity, their use poses non–trivial hindrances to the inferential procedure, particularly from a computational perspective and in high–dimensional scenarios. This still motivates thriving active research for probit, logit, and a number of their generalizations, especially within the Bayesian community. Conjugacy results for standard probit regression under normal and unified skew–normal (SUN) priors appeared only recently in the literature. Such findings were rapidly extended to different generalizations of probit regression, including multinomial probit, dynamic multivariate probit and skewed Gaussian processes among others. Nonetheless, these recent developments focus on specific subclasses of models, which can all be regarded as instances of a potentially broader family of formulations, that rely on partially or fully discretized Gaussian latent utilities. As such, we develop a unified comprehensive framework that encompasses all the above constructions and many others, such as tobit regression and its extensions, for which conjugacy results are yet missing. We show that the SUN family of distribution is conjugate for all models within the broad class considered, which notably encompasses all formulations relying on likelihoods given by the product of multivariate Gaussian densities and cumulative distributions, evaluated at a linear combination of the parameter of interest. Such a unifying framework is practically and conceptually useful for studying general theoretical properties and developing future extensions. This includes new avenues for improved posterior inference exploiting i.i.d. samplers from the exact SUN posteriors and recent accurate and scalable variational Bayes (VB) approximations and expectation–propagation, for which we derive a novel efficient implementation.

Along a parallel research line, we focus on binary regression under logit mapping, for which computations in high dimensions still pose open challenges. To overcome such difficulties, several contributions focus on solving iteratively a series of surrogate problems, entailing the sequential refinement of tangent lower bounds for the logistic log-likelihoods. For instance, tractable quadratic minorizers can be exploited to obtain maximum likelihood (ML) and maximum a posteriori estimates via minorize–maximize and expectation–maximization schemes, with desirable convergence guarantees. Likewise, quadratic surrogates can be used to construct Gaussian approximations of the posterior distribution in mean-field VB routines, which might however suffer from low accuracy in high dimen-

sions. This issue can be mitigated by resorting to more flexible but involved piece-wise quadratic bounds, that however are typically defined in an implicit way and entail reduced tractability as the number of pieces increases. For this reason, we derive a novel tangent minorizer for logistic log-likelihoods, that combines the quadratic term with a single piece-wise linear contribution per each observation, proportional to the absolute value of the corresponding linear predictor. The proposed bound is guaranteed to improve the accuracy over the sharpest among quadratic minorizers, while minimizing the reduction in tractability compared to general piece-wise quadratic bounds. As opposed to the latter, its explicit analytical expression allows to simplify computations by exploiting a renowned scale-mixture representation of Laplace random variables. We investigate the benefit of the proposed methodology both in the context of penalized ML estimation, where it leads to a faster convergence rate of the optimization procedure, and of VB approximation, as the resulting accuracy improvement over mean-field strategies can be substantial in skewed and high-dimensional scenarios.

Contents

Introduction

Binary and categorical regression problems play undoubtedly a pivotal role within the statistics literature (Agresti, 2013). While being of paramount interest per se in a number of applied fields, ranging from social sciences to biostatistics and econometrics, several models developed for binary regression can be exploited as building blocks in more complex statistical constructions, such as density regression (Rodríguez & Dunson, 2011) and additive trees (Chipman et al., 2010). Two of the most common approaches to model discrete observations are arguably given by logistic and probit regression, for which a large and well-established literature is available. Despite that, binary and categorical regression in high-dimensional regimes under probit and logit links still presents several open challenges that motivate thriving active research, primarily among the Bayesian community (Chopin & Ridgway, 2017; Durante, 2019; Fasano et al., 2022) and, to a lesser extent, even in the Frequentist one (Sur & Candès, 2019; Candès & Sur, 2020).

For instance, conjugacy in probit models under the commonly assumed normal prior was believed to be obstructed by the product of Gaussian cumulative distribution functions appearing in the likelihoods. Nonetheless, recent results by Durante (2019) showed that the posterior actually belongs to a known family of distribution, namely the unified skew–normals (Arellano-Valle & Azzalini, 2006; Azzalini & Capitanio, 2013), which generalizes the multivariate Gaussian by introducing skewness while retaining several of its desirable properties. Along the lines of the original result by Durante (2019), in the first Chapter of the present Thesis we focus on Bayesian inference for a large class of models that share the same analytical and computational difficulties of probit regression. Indeed a broad class of models that routinely appear in several fields can be expressed as partially or fully discretized Gaussian linear regressions (Greene, 2008). Besides including classical Gaussian response settings, this class also encompasses probit (Albert & Chib, 1993; Holmes & Held, 2006; Chopin & Ridgway, 2017), multinomial probit (McCulloch & Rossi, 1994; Albert & Chib, 2001; Imai & Van Dyk, 2005) and tobit regression (Tobin, 1958; Chib et al., 2009; Loaiza-Maya et al., 2021), among others, thereby yielding to one of the most widely–implemented families of models in routine applications. The relevance of such representations has stimulated decades of research in the Bayesian field (Albert & Chib, 1993; McCulloch et al., 2000; Albert & Chib, 2001; Andrieu & Doucet, 2002; Girolami & Rogers, 2006; Consonni & Marin, 2007), mostly motivated by the fact that, unlike for the Gaussian linear regression, the posterior distribution induced by such models does not

apparently belong to a known class, under the commonly-assumed Gaussian priors for the coefficients. This has motivated several solutions for posterior inference relying either on sampling–based strategies (Holmes & Held, 2006) or on deterministic approximations (Chopin & Ridgway, 2017) that, however, still experience computational and accuracy issues, especially in high dimensions. The scope of the first Chapter of the present Thesis is to review, unify and extend recent advances in Bayesian inference and computation for this core class of models. To address such a goal, we prove that the likelihoods induced by these formulations share a common analytical structure that implies conjugacy with a broad class of distributions, namely the unified skew–normals, that generalize Gaussians to skewed contexts. This result unifies and extends recent conjugacy properties for specific models within the class analyzed (Durante, 2019; Fasano et al., 2021; Fasano & Durante, 2022), and opens new avenues for improved posterior inference, under a broader class of formulations and priors, via novel closed–form expressions, i.i.d.samplers from the exact SUN posteriors, and more accurate and scalable approximations from variational Bayes (Fasano & Rebaudo, 2021; Blei et al., 2017) and expectation-propagation (Minka, 2001; Vehtari et al., 2020). In particular, we develop a scalable implementation of expectation-propagation with linear cost per iteration in the number of covariates, as opposed to the quadratic cost achieved by state-of-the-art implementations that greatly hampers their use in high-dimensional settings. All aforementioned advantages are illustrated in simulations and are expected to facilitate the routine–use of these core Bayesian models, while providing novel frameworks for studying general theoretical properties and developing future extensions.

Along a parallel line of research, in the second Chapter of the present Thesis we turn the attention to binary regression under logit mapping, which notoriously hinders tractable analytical inference. In the attempt to circumvent such difficulty, data-augmentation strategies for logistic regression have received considerable attention within the Bayesian framework (Holmes & Held, 2006; Frühwirth-Schnatter & Frühwirth, 2007; Frühwirth-Schnatter et al., 2009; Zens et al., 2020; Polson et al., 2012). Conversely, unconstrained and penalized maximum likelihood estimations typically proceed via iterative schemes that alternate between the construction and the optimization of quadratic approximations of the logistic log-likelihood (Böhning & Lindsay, 1988; Jaakkola & Jordan, 2000), either corresponding to Newton's method or arising from different tangent bounds exploited within minorize-maximize (Hunter & Lange, 2004; Wu & Lange, 2010) or expectation-maximization schemes (Dempster et al., 1977; McLachlan & Krishnan, 1996). As Newton's method remains prone to unstable convergence issues, we focus our attention on the above strategies, leveraging on the optimality of the lower bound corresponding to the Pólya-Gamma data-augmentation scheme (Polson et al., 2012; Durante & Rigon, 2019) among quadratic minorizers for the logistic log-likelihood. We highlight how such advantage over alternative quadratic bounds is enhanced by the combination with ℓ_1-regularizations, such as lasso (Tibshirani, 1996) or elastic net (Zou & Hastie, 2005). Indeed, the presence of the non-smooth penalty

contribution dictates the use of coordinate-wise optimization schemes (Friedman et al., 2007; Hastie et al., 2015), that indirectly levels out the difference in computational cost that arises from expensive algebraic operations in the unpenalized case (Durante & Rigon, 2019). Furthermore, we derive a novel tangent minorizer, dominating the Pólya-Gamma one, by adding a piece-wise linear contribution proportional to the ℓ_1-norm of the linear predictors. Strictly speaking, the proposed methodology falls within the general class of piece-wise quadratic tangent bounds for logistic likelihoods, such as those proposed by Khan et al. (2010) and Marlin et al. (2011), which are known to improve over the traditional quadratic bounds by Böhning & Lindsay (1988) and Jaakkola & Jordan (2000). However, their construction remains inherently implicit, and the actual bound is found by solving numerically a minimax problem, agnostic to the observed data, which requires imposing an arbitrary number of piece-wise quadratic branches for each approximate likelihood term. On the contrary, the novel bound we propose allows for an explicit analytical representation, parametrized by a set of variational locations which are learned by the data as part of the inferential procedure. Notably, such piece-wise quadratic bound still allows for exact solutions of the corresponding coordinate-wise optimization equations. Empirical results support the intuition that the higher flexibility of the proposed bound lead to an improved convergence rate of the resulting minorize-maximize scheme.

Finally, we further leverage on the novel piece-wise quadratic bound to construct an improved variational approximation of the posterior distribution (Blei et al., 2017; Ormerod & Wand, 2010). Indeed variational Bayes routines provide a popular class of strategies to perform approximate posterior computations, whenever a faster alternative to sampling-based inference is required. The essence of such methods lies in the minimization of a suitable discrepancy, most often the forward Kullback-Leiber (Kullback & Leibler, 1951) divergence, between the exact posterior and an approximate one, belonging to a pre-specified family of distributions. The latter is typically identified by the enforcement of an explicit analytical form or via the imposition of a specific dependence structure in the target parameter space (Bishop, 2006). Either way, the choice of approximating class is driven by an implicit trade-off between tractability, which ensures the computational advantage over sampling schemes, and flexibility, which eventually allows for accurate approximation. In the case of Bayesian logistic regression, several contributions in the literature resort to a Gaussian approximation for the intractable posterior, originally derived in Jaakkola & Jordan (2000) by direct tangent minorization of the log-likelihood contributions. Only recently, Durante & Rigon (2019) showed that such procedure is actually endowed with a full probabilistic interpretation, as it arises as a proper mean-field variational Bayes routine (Blei et al., 2017) under the celebrated Pólya-Gamma data augmentation scheme by Polson et al. (2012). In the third Chapter we exploit the piece-wise linear-quadratic bound derived previously to construct a more accurate variational approximation of the posterior in logistic regression models, which dominates over the Pólya-Gamma mean-field one as a consequence of relative tightness of the corresponding log-likelihood lower bounds. Intu-

itively, the advantage of the novel approximation comes from a higher flexibility in choosing the posterior location and inflating its variance by tuning a set of associated variational parameters. The novel approximate procedure still allows for simple expression of the update equations for such variational parameters, albeit entailing the evaluation of suitable expected values with respect to a distribution with piece-wise quadratic kernel. Nonetheless, it is possible to obviate the lack of closed-form expressions by exploiting a well-known scale mixture of normals representation for the Laplace contributions appearing in the lower bound, previously exploited in the literature dealing with the Bayesian versions of lasso (Park & Casella, 2008; Hans, 2009) or quantile regression (Kozumi & Kobayashi, 2011; Li et al., 2010). This enables to obtain Monte Carlo estimates of the desired quantity, notably allowing for an implementation with linear cost in the number of covariates, given each sample of the additional auxiliary variables arising from the scale mixture representation. As a consequence, the resulting approximation still leads to a positive tractability-accuracy trade-off in large-p-small-n scenarios, where state-of-the-art exact sampling schemes often face severe limitations while mean-field variational Bayes might suffer from reduced accuracy. Finally, a reverse engineering process on the piece-wise linear-quadratic bound might lead to the construction of a novel data augmentation scheme, which would allow for a fully probabilistic interpretation of the proposed variational procedure.

Chapter 1

Bayesian conjugacy in probit, tobit, multinomial probit and extensions: a review and new results

A broad class of models that routinely appear in several fields can be expressed as partially or fully discretized Gaussian linear regressions. Besides including basic Gaussian response settings, this class also encompasses probit, multinomial probit and tobit regression, among others, thereby yielding one of the most widely–implemented families of models in routine applications. The relevance of such representations has stimulated decades of research in the Bayesian field, mostly motivated by the fact that, unlike for the Gaussian linear regression, the posterior distribution induced by such models does not apparently belong to a known class, under the commonly–assumed Gaussian priors for the coefficients. This has motivated several solutions for posterior inference relying either on sampling–based strategies or on deterministic approximations that, however, still experience computational and accuracy issues, especially in high dimensions. The scope of this Chapter is to review, unify and extend recent advances in Bayesian inference and computation for this core class of models. To address such a goal, we prove that the likelihoods induced by these formulations share a common analytical structure that implies conjugacy with a broad class of distributions, namely the unified skew–normals, that generalize Gaussians to skewed contexts. This result unifies and extends recent conjugacy properties for specific models within the class analyzed, and opens new avenues for improved posterior inference, under a broader class of formulations and priors, via novel closed–form expressions, i.i.d. samplers from the exact SUN posteriors, and more accurate and scalable approximations from variational Bayes and expectation–propagation. Such advantages are illustrated in simulations and are expected to facilitate the routine–use of these core Bayesian models, while providing a novel framework for studying general theoretical properties and developing future extensions.

1.1 Introduction

The scope of this Chapter is to review, unify, compare and extend both past and recent developments in Bayesian inference for probit (Bliss, 1934), multinomial probit (Hausman & Wise, 1978; Tutz, 1991; Stern, 1992) and tobit (Tobin, 1958) models, along with their extensions to multivariate, skewed, non–linear and dynamic contexts. Although such models provide core formulations in statistics (DeMaris, 2004; Greene, 2008; Agresti, 2013) and often appear as building–blocks within more complex constructions (see e.g., Chipman et al., 2010; Rodríguez & Dunson, 2011), Bayesian inference under the associated likelihoods still presents open challenges that have motivated decades of active research in the field (Chopin & Ridgway, 2017). This is mainly due to the presence in the likelihood of Gaussian cumulative distribution functions arising from a partial or full discretization of a set of Gaussian latent utilities under a discrete choice perspective (e.g., Greene, 2008), which clearly hinders conjugacy when combined with the common Gaussian priors for the regression coefficients $\boldsymbol{\beta}$.

This lack of conjugacy for such a broad and routinely–used class of models motivates still ongoing efforts to develop effective Markov Chain Monte Carlo (MCMC)–based sampling methods and accurate deterministic approximations of the posterior distribution to perform Bayesian inference in probit (Albert & Chib, 1993; Holmes & Held, 2006; Consonni & Marin, 2007; Pakman & Paninski, 2014; Chopin & Ridgway, 2017), tobit (Chib, 1992; Chib et al., 2009; Loaiza-Maya et al., 2021), multinomial probit (Albert & Chib, 1993; McCulloch & Rossi, 1994; Nobile, 1998; McCulloch et al., 2000; Albert & Chib, 2001; Imai & Van Dyk, 2005) and their extensions to multivariate, skewed, dynamic and non–linear settings (Chib & Greenberg, 1998; Chen et al., 1999; Andrieu & Doucet, 2002; Sahu et al., 2003; Kuss et al., 2005; Girolami & Rogers, 2006; Bazán et al., 2010; Talhouk et al., 2012; Soyer & Sung, 2013; Riihimäki et al., 2014). Although these methods yield state–of–the–art implementations, there are still key open questions on computational scalability, mixing and approximation accuracy, especially in high dimensions (Chopin & Ridgway, 2017). These issues, combined with the recent conjugacy results for probit models in Durante (2019), have led to renewed interest in closed–form solutions for Bayesian inference under these formulations. More specifically, Durante (2019) recently proved that the posterior distribution for the $\boldsymbol{\beta}$ coefficients in Bayesian probit regression under Gaussian priors belongs to the class of unified skew–normals (Arellano-Valle & Azzalini, 2006; Azzalini & Capitanio, 2013) and, more generally, that SUNs are conjugate to probit regression models. The SUN class extends multivariate Gaussians to include skewness, and its analytical properties have led to rapid subsequent extensions of the original results to multinomial probit (Fasano & Durante, 2022), dynamic multivariate probit (Fasano et al., 2021), Gaussian processes (Cao et al., 2022), skewed Gaussian processes (Benavoli et al., 2020, 2021), skew–elliptical link functions (Zhang et al., 2021a) and rounded data (Kowal, 2021), while facilitating the development of improved approximations (Fasano et al., 2022; Fasano &

Durante, 2022; Fasano & Rebaudo, 2021).

The above advancements are providing yet unexplored opportunities for Bayesian inference under such models via novel closed–form expressions, tractable Monte Carlo methods relying on i.i.d. samples from the exact SUN posteriors, and more accurate and scalable approximations from variational Bayes (e.g., Blei et al., 2017) and expectation–propagation (EP) (e.g., Chopin & Ridgway, 2017). However, most of these new developments focus on specific sub–classes of models within a potentially broader family of formulations that rely on partially or fully discretized Gaussian latent utilities. Therefore, there is still the lack of a unified framework that would be practically and conceptually useful to derive general conjugacy results along with broadly–applicable closed–form solutions, Monte Carlo methods and improved approximations of the posterior distribution. For instance, conjugacy results for tobit models (Tobin, 1958) are yet missing in the literature, however, as it will be clarified in Section 1.3, SUNs are conjugate also to this class. Such a comprehensive treatment would also help to clarify these advancements in the light of previously–developed state–of–the–art MCMC methods and approximations, and would serve as a catalyst of applied, methodological and theoretical research to further expand the set of solutions for this broad class of models.

This Chapter aims at addressing the above gap to boost the routine–use of these core Bayesian models, and provide comprehensive frameworks for studying general theoretical properties and developing future extensions. As a first step toward accomplishing this goal, Section 1.2 unifies probit, tobit, multinomial probit and related extensions by reformulating the associated likelihoods as special cases of a general form that relies on the product of multivariate Gaussian densities and cumulative distributions, both evaluated at a linear combination of the coefficients $\boldsymbol{\beta}$. Such a unified formulation is crucial to prove a general result in Section 1.3 which states that SUN distributions are conjugate priors to any model whose likelihood can be expressed as a special case of the one defined in Section 1.2. This result unifies available findings for probit (Durante, 2019), multinomial probit (Fasano & Durante, 2022) and dynamic multivariate probit (Fasano et al., 2021), among others, while extending SUN conjugacy properties to a much broader class of Bayesian models for which similar results have not appeared yet in the literature. Notable examples are tobit models (Tobin, 1958) and any extension of probit, tobit and multinomial probit which replaces the Gaussian latent utilities with skew–normal ones (Chen et al., 1999; Sahu et al., 2003; Bazán et al., 2010), among others. As discussed in Section 1.4, this unified conjugacy result is also practically–relevant since it allows to inherit all the recent methodological and computational developments for Bayesian inference under SUN posteriors in probit and multinomial probit to the whole class of models presented in Section 1.2. Such advancements include novel closed–form expressions for relevant posterior moments, marginal likelihoods and predictive distributions, along with improved Monte Carlo methods and deterministic approximations from variational Bayes and expectation–propagation. These solutions are presented in detail in Section 1.4 along with a careful review previous state–

of–the–art solutions, recast under the proposed general framework. An excellent review of these previous solutions can be already found in Chopin & Ridgway (2017), but the focus is on univariate probit models. Due to this, the present Chapter will mostly consider the more recent developments relying on SUN conjugacy and on their discussion in the light of previous solutions, when adapted to the broader class of models and priors, beyond classical Bayesian probit regression. Consistent with this scope, Section 1.6 concludes with a general discussion that points toward several future research directions motivated by the unified framework developed in this Chapter. Empirical studies illustrating the potential of this unification are provided in Section 1.5.

1.2 A unified likelihood for probit, tobit, multinomial probit and extensions

As discussed in the previos Section, probit (Bliss, 1934), tobit (Tobin, 1958), multinomial probit (Hausman & Wise, 1978; Tutz, 1991; Stern, 1992) and their extensions are core formulations in statistics, and, when seen as specific examples of a more general representation which also includes classical Gaussian linear regression, arguably yield one of the most widely–implemented classes of models in routine applications (DeMaris, 2004; Greene, 2008; Agresti, 2013). In fact, all these formulations share a common generative construction, in that the corresponding responses can be defined as partially or fully discretized versions of continuous ones from a set of underlying Gaussian linear regressions (Chib, 1992; Albert & Chib, 1993; Chib & Greenberg, 1998). More specifically, let $z_i \in \Re$ denote a latent continuous response available for every unit $i = 1, \dots, n$, and consider the standard linear regression model $z_i = \mathbf{x}_i^\intercal \boldsymbol{\beta} + \varepsilon_i$, with noise $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, covariates' vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\intercal$ and coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\intercal$. Starting from this building–block formulation, classical Gaussian linear regression models, probit models (Bliss, 1934) and tobit regression (Tobin, 1958) can be obtained by letting $y_i = z_i$, $y_i = \mathbb{1}(z_i > 0)$ and $y_i = \max\{z_i, 0\} = z_i \mathbb{1}(z_i > 0)$, respectively. The first two constructions correspond to the limiting cases in which z_i is either entirely observed or dichotomized, respectively, whereas the third one represents the intermediate situation in which z_i is fully observed only if it exceeds value 0 (Chib, 1992; Albert & Chib, 1993). Multinomial probit models (Hausman & Wise, 1978; Tutz, 1991; Stern, 1992) for categorical responses $y_i \in \{1; \dots; L\}$ can be derived with a similar reasoning. For instance, in the formulation proposed by Stern (1992), the observed categorical response y_i is defined as $y_i = \mathrm{argmax}_l\{z_{i1}, \dots, z_{iL}\}$ where $z_{i1}, \dots, z_{iL}$ are class–specific Gaussian latent utilities related to the covariates via a system of linear regressions $z_{il} = \mathbf{x}_i^\intercal \boldsymbol{\beta}_l + \varepsilon_{il}$ for each $l = 1, \dots, L$, with $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{iL})^\intercal \sim \mathrm{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$.

As shown in Sections 1.2.1–1.2.4, these similarities in the generative models imply that the likelihoods induced by the above formulations and their extensions are all specific examples of the general form

$$p(\mathbf{y} \mid \boldsymbol{\beta}) = p(\bar{\mathbf{y}}_1 \mid \boldsymbol{\beta}) p(\bar{\mathbf{y}}_0 \mid \boldsymbol{\beta}) \propto \phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_1)\Phi_{\bar{n}_0}(\bar{\mathbf{y}}_0 + \bar{\mathbf{X}}_0\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_0), \tag{1.1}$$

with $\phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_1)$ and $\Phi_{\bar{n}_0}(\bar{\mathbf{y}}_0 + \bar{\mathbf{X}}_0\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_0)$ denoting the density and the cumulative distribution function of the multivariate Gaussians $\mathrm{N}_{\bar{n}_1}(\mathbf{0}, \bar{\boldsymbol{\Sigma}}_1)$ and $\mathrm{N}_{\bar{n}_0}(\mathbf{0}, \bar{\boldsymbol{\Sigma}}_0)$, evaluated at $\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\beta}$ and $\bar{\mathbf{y}}_0 + \bar{\mathbf{X}}_0\boldsymbol{\beta}$, respectively, where $\bar{\mathbf{y}}_1 := \bar{\mathbf{y}}_1(\mathbf{y})$ and $\bar{\mathbf{y}}_0 := \bar{\mathbf{y}}_0(\mathbf{y})$ denote known response vectors obtained as a function of $\mathbf{y}$, whereas $\bar{\mathbf{X}}_1 := \bar{\mathbf{X}}_1(\mathbf{y}, \mathbf{X})$ and $\bar{\mathbf{X}}_0 := \bar{\mathbf{X}}_0(\mathbf{y}, \mathbf{X})$ are suitable design matrices which can be directly derived from the observed predictors in $\mathbf{X}$ and, possibly, the response vector $\mathbf{y}$. For instance, the likelihood $\prod_{i=1}^n \Phi(\mathbf{x}_i^\intercal\boldsymbol{\beta})^{\mathbb{1}(y_i=1)}[1 - \Phi(\mathbf{x}_i^\intercal\boldsymbol{\beta})]^{\mathbb{1}(y_i=0)}$ under classical probit regression can be rewritten as $\prod_{i=1}^n \Phi[(2y_i - 1)\mathbf{x}_i^\intercal\boldsymbol{\beta}] = \Phi_n[\mathrm{diag}(2\mathbf{y} - \mathbf{1}_n)\mathbf{X}\boldsymbol{\beta}; \mathbf{I}_n]$, which coincides with (1.1) after letting $\bar{n}_1 = 0$, $\bar{n}_0 = n$, $\bar{\mathbf{y}}_0 = \mathbf{0}$, $\bar{\mathbf{X}}_0 = \mathrm{diag}(2\mathbf{y} - \mathbf{1}_n)\mathbf{X}$ and $\bar{\boldsymbol{\Sigma}}_0 = \mathbf{I}_n$. As shown in Sections 1.2.1–1.2.4, similar results can be obtained under tobit, multinomial probit and other relevant extensions of these models.

1.2.1 Linear regression, multivariate linear regression and extensions

Although the focus of this Chapter is on models beyond the classical Gaussian response setting, it is worth emphasizing that also this class induces likelihoods which are special cases of (1.1). For instance, standard Gaussian linear regression $(y_i \mid \boldsymbol{\beta}) \sim \mathrm{N}(\mathbf{x}_i^\intercal\boldsymbol{\beta}, \sigma^2)$, independently for $i = 1, \ldots, n$, is directly recovered after noticing that the induced likelihood

$$p(\mathbf{y} \mid \boldsymbol{\beta}) \propto \prod_{i=1}^n \phi(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}; \sigma^2) = \phi_n(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}; \sigma^2\mathbf{I}_n), \tag{1.2}$$

coincides with (1.1), when letting $\bar{n}_0 = 0$, $\bar{n}_1 = n$, $\bar{\mathbf{y}}_1 = \mathbf{y}$, $\bar{\mathbf{X}}_1 = \mathbf{X}$ and $\bar{\boldsymbol{\Sigma}}_1 = \sigma^2\mathbf{I}_n$. As a direct consequence, also heteroscedastic and correlated versions can be incorporated by replacing $\sigma^2\mathbf{I}_n$ with a general residuals covariance matrix. Similarly, the likelihood associated with multivariate Gaussian response data from the regression model $(\mathbf{y}_i \mid \boldsymbol{\beta}) \sim \mathrm{N}_m(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$, independently for $i = 1, \ldots, n$, can be written as

$$p(\mathbf{y} \mid \boldsymbol{\beta}) \propto \prod_{i=1}^n \phi_m(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}; \boldsymbol{\Sigma}) = \phi_{n\cdot m}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}; \mathbf{I}_n \otimes \boldsymbol{\Sigma}), \tag{1.3}$$

where $\mathbf{y} = (\mathbf{y}_1^\intercal, \ldots, \mathbf{y}_n^\intercal)^\intercal$, $\mathbf{X} = (\mathbf{X}_1^\intercal, \ldots, \mathbf{X}_n^\intercal)^\intercal$ and $\otimes$ denotes the Kronecker product. Setting $\bar{n}_0 = 0$, $\bar{n}_1 = n \cdot m$, $\bar{\mathbf{y}}_1 = \mathbf{y}$, $\bar{\mathbf{X}}_1 = \mathbf{X}$ and $\bar{\boldsymbol{\Sigma}}_1 = \mathbf{I}_n \otimes \boldsymbol{\Sigma}$ in (1.1) yields directly to (1.3).

Eventually, it is also possible to include skewness in the above formulation, while still remaining within the class of models whose likelihood can be expressed as in (1.1). Recalling Sahu et al. (2003) and Azzalini (2005), this can be done by assuming that $(y_i \mid \boldsymbol{\beta}) \sim \mathrm{SN}(\mathbf{x}_i^\intercal\boldsymbol{\beta}, \sigma^2, \alpha)$, independently for $i = 1, \ldots, n$, where $\mathrm{SN}(\mathbf{x}_i^\intercal\boldsymbol{\beta}, \sigma^2, \alpha)$ denotes the skew–normal distribution (Azzalini, 1985) with location $\mathbf{x}_i^\intercal\boldsymbol{\beta}$, scale σ^2 and shape parameter α. This choice implies that

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\beta}) &\propto \prod_{i=1}^n \phi(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}; \sigma^2)\Phi(\alpha(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}); \sigma^2) \\ &= \phi_n(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}; \sigma^2\mathbf{I}_n)\Phi_n(\alpha\mathbf{y} - \alpha\mathbf{X}\boldsymbol{\beta}; \sigma^2\mathbf{I}_n), \end{aligned} \tag{1.4}$$

which coincides again with equation (1.1) when $\bar{n}_1 = \bar{n}_0 = n$, $\bar{\mathbf{y}}_1 = \mathbf{y}$, $\bar{\mathbf{y}}_0 = \alpha\mathbf{y}$, $\bar{\mathbf{X}}_1 = \mathbf{X}$, $\bar{\mathbf{X}}_0 = -\alpha\mathbf{X}$ and $\bar{\boldsymbol{\Sigma}}_1 = \bar{\boldsymbol{\Sigma}}_0 = \sigma^2\mathbf{I}_n$. Inclusion of skewed responses from more elaborated distributions such as the multivariate skew–normal (Azzalini & Dalla Valle, 1996; Azzalini & Capitanio, 1999), the extended multivariate skew–normal (Arnold & Beaver, 2000; Arnold et al., 2002), the closed skew–normal family (González-Farias et al., 2004; Gupta et al., 2004) and the SUN (Arellano-Valle & Azzalini, 2006), is also possible and yields again special cases of (1.1), with similar derivations.

1.2.2 Probit, multivariate probit, multinomial probit and extensions

As discussed in Section 1.2, the classical probit regression model $(y_i \mid \boldsymbol{\beta}) \sim \text{Bern}[\Phi(\mathbf{x}_i^\intercal\boldsymbol{\beta})]$, independently for $i = 1, \ldots, n$, induces likelihoods that can be readily reframed within representation (1.1). More specifically, recalling Durante (2019), the probit likelihood can be expressed as

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\beta}) \propto \prod_{i=1}^n \Phi(\mathbf{x}_i^\intercal\boldsymbol{\beta})^{y_i}[1-\Phi(\mathbf{x}_i^\intercal\boldsymbol{\beta})]^{1-y_i} &= \prod_{i=1}^n \Phi[(2y_i-1)\mathbf{x}_i^\intercal\boldsymbol{\beta}] \\ &= \Phi_n(\text{diag}(2\mathbf{y}-\mathbf{1}_n)\mathbf{X}\boldsymbol{\beta}; \mathbf{I}_n), \end{aligned} \tag{1.5}$$

which is a special case of equation (1.1), after letting $\bar{n}_1 = 0$, $\bar{n}_0 = n$, $\bar{\mathbf{y}}_0 = \mathbf{0}$, $\bar{\mathbf{X}}_0 = \text{diag}(2\mathbf{y}-\mathbf{1}_n)\mathbf{X}$ and $\bar{\boldsymbol{\Sigma}}_0 = \mathbf{I}_n$. Probit probabilities of the form $\Phi(\mathbf{x}_i^\intercal\boldsymbol{\beta}; \sigma^2)$ can also be included by replacing $\bar{\boldsymbol{\Sigma}}_0 = \mathbf{I}_n$ with $\bar{\boldsymbol{\Sigma}}_0 = \sigma^2\mathbf{I}_n$.

The above probit model also admits a number of routinely–used extensions to incorporate multivariate binary outcomes (Chib & Greenberg, 1998) and multinomial response data (Hausman & Wise, 1978; Tutz, 1991; Stern, 1992). As previously mentioned, both cases have their roots in discrete choice models (e.g., Greene, 2008), and can be reframed within (1.1). To clarify this result, let us first focus on multivariate probit models for the binary response vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^\intercal \in \{0; 1\}^m$. As discussed in Chib & Greenberg (1998), these formulations can be interpreted as a dichotomized version of the regression model for multivariate Gaussian response data in Section 1.2.1. In fact, each $\mathbf{y}_i$ is defined as $\mathbf{y}_i = [\mathbb{1}(z_{i1} > 0), \ldots, \mathbb{1}(z_{im} > 0)]^\intercal$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{im})^\intercal \sim \text{N}_m(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$, independently for every $i = 1, \ldots, n$. This means that the contribution to the likelihood of each unit i is $p(\mathbf{y}_i \mid \boldsymbol{\beta}) = p([\mathbb{1}(z_{i1} > 0), \ldots, \mathbb{1}(z_{im} > 0)]^\intercal \mid \boldsymbol{\beta})$, which can be also written as $\Phi_m(\mathbf{B}_i\mathbf{X}_i\boldsymbol{\beta}; \mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_i)$, following standard properties of multivariate Gaussian cumulative distribution function, where $\mathbf{B}_i = \text{diag}(2y_{i1} - 1, \ldots, 2y_{im} - 1)$. As a result, the joint likelihood of multivariate probit regression is

$$p(\mathbf{y} \mid \boldsymbol{\beta}) \propto \prod_{i=1}^n \Phi_m(\mathbf{B}_i\mathbf{X}_i\boldsymbol{\beta}; \mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_i) = \Phi_{n\cdot m}(\mathbf{B}\mathbf{X}\boldsymbol{\beta}; \mathbf{B}(\mathbf{I}_n \otimes \boldsymbol{\Sigma})\mathbf{B}), \tag{1.6}$$

where $\mathbf{X} = (\mathbf{X}_1^\intercal, \ldots, \mathbf{X}_n^\intercal)^\intercal$, and $\mathbf{B}$ denotes a block–diagonal matrix with generic block $\mathbf{B}_{[i,i]} = \mathbf{B}_i$, for each $i = 1, \ldots, n$. To reframe (1.6) within the general likelihood form in (1.1) it suffices to set $\bar{n}_1 = 0$, $\bar{n}_0 = n \cdot m$, $\bar{\mathbf{y}}_0 = \mathbf{0}$, $\bar{\mathbf{X}}_0 = \mathbf{B}\mathbf{X}$ and $\bar{\boldsymbol{\Sigma}}_0 = \mathbf{B}(\mathbf{I}_n \otimes \boldsymbol{\Sigma})\mathbf{B}$.

As discussed in Fasano & Durante (2022), similar constructions and derivations can be also considered for a variety of multinomial probit models (Hausman & Wise, 1978; Tutz, 1991; Stern, 1992). All these formulations express the probabilities of the L different categories $\{1; \ldots; L\}$ via a discrete choice mechanism relying on correlated predictor–dependent Gaussian latent utilities which facilitate improved flexibility and avoid restrictive assumptions often found in multinomial logit, such as the independence of irrelevant alternatives (Hausman & Wise, 1978). For instance, in the formulation by Stern (1992), each categorical response variable y_i is defined as $y_i = \text{argmax}_l\{z_{i1}, \ldots, z_{iL}\}$, where $z_{il} = \mathbf{x}_i^\intercal \boldsymbol{\beta}_l + \varepsilon_{il}$ for $l = 1, \ldots, L$, with $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iL})^\intercal \sim \text{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\beta}_L = \mathbf{0}$ for identifiability purposes (Johndrow et al., 2013). As a direct consequence of this construction, it follows that $\text{pr}(y_i = l \mid \boldsymbol{\beta}) = \text{pr}(z_{il} > z_{ik}, \text{ for each } k \neq l) = \text{pr}(\varepsilon_{ik} - \varepsilon_{il} < \mathbf{x}_i^\intercal \boldsymbol{\beta}_l - \mathbf{x}_i^\intercal \boldsymbol{\beta}_k, \text{ for each } k \neq l)$. Therefore, let $\mathbf{v}_l$ denote the $L \times 1$ vector with value 1 in position l and 0 elsewhere, for every $l = 1, \ldots, L$, and define $\mathbf{x}_{il} = \bar{\mathbf{v}}_l \otimes \mathbf{x}_i$, where $\bar{\mathbf{v}}_l$ is the $(L-1) \times 1$ vector obtained by removing the L–th entry from $\mathbf{v}_l$. Then, $\text{pr}(y_i = l \mid \boldsymbol{\beta}) = \text{pr}[(\mathbf{v}_k - \mathbf{v}_l)^\intercal \boldsymbol{\varepsilon}_i < (\mathbf{x}_{il} - \mathbf{x}_{ik})^\intercal \boldsymbol{\beta}, \text{ for each } k \neq l]$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_{L-1}^\intercal)^\intercal$. This expression for the probability of a generic category l can be also formulated in the more compact form $\text{pr}(\mathbf{V}_{[-l]}\boldsymbol{\varepsilon}_i < \mathbf{X}_{i[-l]}\boldsymbol{\beta})$, where $\mathbf{V}_{[-l]}$ and $\mathbf{X}_{i[-l]}$ denote suitable matrices whose rows are obtained by stacking the vectors $(\mathbf{v}_k - \mathbf{v}_l)^\intercal$ and $(\mathbf{x}_{il} - \mathbf{x}_{ik})^\intercal$, respectively, for every $k \neq l$. Therefore, leveraging the standard properties of multivariate Gaussians, as done for the multivariate probit setting, it follows that $\text{pr}(y_i = l \mid \boldsymbol{\beta}) = \text{pr}(\mathbf{V}_{[-l]}\boldsymbol{\varepsilon}_i < \mathbf{X}_{i[-l]}\boldsymbol{\beta}) = \Phi_{L-1}(\mathbf{X}_{i[-l]}\boldsymbol{\beta}; \mathbf{V}_{[-l]}\boldsymbol{\Sigma}\mathbf{V}_{[-l]}^\intercal)$, for every $l = 1, \ldots, L$. This result yields a joint likelihood for the observed categorical responses $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ which can be written as

$$p(\mathbf{y} \mid \boldsymbol{\beta}) \propto \prod_{i=1}^{n} \Phi_{L-1}(\mathbf{X}_{i[-y_i]}\boldsymbol{\beta}; \mathbf{V}_{[-y_i]}\boldsymbol{\Sigma}\mathbf{V}_{[-y_i]}^\intercal) = \Phi_{n\cdot(L-1)}(\mathbf{X}\boldsymbol{\beta}; \mathbf{V}(\mathbf{I}_n \otimes \boldsymbol{\Sigma})\mathbf{V}^\intercal), \tag{1.7}$$

where $\mathbf{X} = (\mathbf{X}_{1[-y_1]}^\intercal, \ldots, \mathbf{X}_{n[-y_n]}^\intercal)^\intercal$, and $\mathbf{V}$ is a block–diagonal matrix with generic block $\mathbf{V}_{[i,i]} = \mathbf{V}_{[-y_i]}$, for each $i = 1, \ldots, n$. Setting $\bar{n}_1 = 0$, $\bar{n}_0 = n \cdot (L-1)$, $\bar{\mathbf{y}}_0 = \mathbf{0}$, $\bar{\mathbf{X}}_0 = \mathbf{X}$ and $\bar{\boldsymbol{\Sigma}}_0 = \mathbf{V}(\mathbf{I}_n \otimes \boldsymbol{\Sigma})\mathbf{V}^\intercal$ in (1.1) leads to equation (1.7). Hence, the multinomial probit model by Stern (1992) is again a special case of the general form in (1.1). As shown in Sections 2.1 and 2.3 of Fasano & Durante (2022), also the alternative formulations proposed by Hausman & Wise (1978) and Tutz (1991) induce likelihoods which can be expressed as cumulative distribution functions of multivariate Gaussians evaluated at a suitable linear combination of the coefficients' vector $\boldsymbol{\beta}$; see Propositions 1 and 3 in Fasano & Durante (2022). This means that also such models can be easily recast within the general form in (1.1) with $\bar{n}_1 = 0$, $\bar{\mathbf{y}}_0 = \mathbf{0}$, and suitably–chosen $\bar{\mathbf{X}}_0$ and $\bar{\boldsymbol{\Sigma}}_0$.

Inclusion of skewness is possible also under probit, multivariate probit and multinomial probit models. This direction has been effectively explored by Chen et al. (1999) and Bazán et al. (2010), with a main focus on basic probit models, and can be again reframed within the general formulation in (1.1). For example, in the context of univariate probit regression, skewness can be incorporated by replacing the Gaussian latent utilities with skew–normal ones; namely $(z_i \mid \boldsymbol{\beta}) \sim \text{SN}(\mathbf{x}_i^\intercal \boldsymbol{\beta}, \sigma^2, \alpha)$, independently for $i = 1, \ldots, n$. As a

consequence, the binary response data $y_i = \mathbb{1}(z_i > 0)$ are Bernoulli variables with parameter $\text{pr}(y_i = 1 \mid \boldsymbol{\beta}) = \text{pr}(z_i > 0 \mid \boldsymbol{\beta}) \propto \Phi_2[(\mathbf{x}_i^\intercal \boldsymbol{\beta}, 0)^\intercal; \text{diag}(\sigma^2, 1) + \sigma\alpha(1+\alpha^2)^{-1/2}(\mathbf{1}_2\mathbf{1}_2^\intercal - \mathbf{I}_2)]$, whose expression follows directly from the the cumulative distribution function of a skew–normal; see e.g., González-Farias et al. (2004); Arellano-Valle & Azzalini (2006); Azzalini & Bacchieri (2010); Azzalini & Capitanio (2013), and Arellano-Valle & Azzalini (2021). Leveraging the same results it also follows that $\text{pr}(y_i = 0 \mid \boldsymbol{\beta}) = \text{pr}(z_i < 0 \mid \boldsymbol{\beta}) \propto \Phi_2[(-\mathbf{x}_i^\intercal \boldsymbol{\beta}, 0)^\intercal; \text{diag}(\sigma^2, 1) - \sigma\alpha(1+\alpha^2)^{-1/2}(\mathbf{1}_2\mathbf{1}_2^\intercal - \mathbf{I}_2)]$. Let $\mathbf{X}_i = [(2y_i - 1)\mathbf{x}_i, \mathbf{0}]^\intercal$ and $\boldsymbol{\Sigma}_i = \text{diag}(\sigma^2, 1) + (2y_i - 1)\sigma\alpha(1+\alpha^2)^{-1/2}(\mathbf{1}_2\mathbf{1}_2^\intercal - \mathbf{I}_2) = \boldsymbol{\Sigma} + (2y_i - 1)\boldsymbol{\Lambda}$, the joint likelihood for the observed binary responses can be then expressed as

$$p(\mathbf{y} \mid \boldsymbol{\beta}) \propto \prod_{i=1}^{n} \Phi_2(\mathbf{X}_i\boldsymbol{\beta}; \boldsymbol{\Sigma}_i) = \Phi_{2n}(\mathbf{X}\boldsymbol{\beta}; \mathbf{I}_n \otimes \boldsymbol{\Sigma} + \text{diag}(2\mathbf{y} - \mathbf{1}_n) \otimes \boldsymbol{\Lambda}), \tag{1.8}$$

where $\mathbf{X} = (\mathbf{X}_1^\intercal, \ldots, \mathbf{X}_n^\intercal)^\intercal$, $\boldsymbol{\Sigma} = \text{diag}(\sigma^2, 1)$, and $\boldsymbol{\Lambda} = \sigma\alpha(1+\alpha^2)^{-1/2}(\mathbf{1}_2\mathbf{1}_2^\intercal - \mathbf{I}_2)$. As a result, equation (1.8) is again a special case of (1.1) after setting $\bar{n}_1 = 0$, $\bar{n}_0 = 2n$, $\bar{\mathbf{y}}_0 = \mathbf{0}$, $\bar{\mathbf{X}}_0 = \mathbf{X}$ and $\bar{\boldsymbol{\Sigma}}_0 = \mathbf{I}_n \otimes \boldsymbol{\Sigma} + \text{diag}(2\mathbf{y} - \mathbf{1}_n) \otimes \boldsymbol{\Lambda}$. Similar derivations can be considered to incorporate skewness in multivariate and multinomial probit via multivariate skew–normal (Azzalini & Dalla Valle, 1996), closed skew–normal (González-Farias et al., 2004; Gupta et al., 2004) or unified skew–normal (Arellano-Valle & Azzalini, 2006) latent utilities. Some of these choices have not yet been explored to induce skewed link functions for multivariate and multinomial extensions of classical probit regression. Nonetheless, all these variables have cumulative distribution functions proportional to those of multivariate Gaussians, evaluated at a linear combination of $\boldsymbol{\beta}$, and, hence, induce likelihoods which can be again expressed as special cases of the general framework in (1.1).

1.2.3 Tobit regression and extensions

Recalling Section 1.2, the classical tobit model (Tobin, 1958) characterizes the intermediate situation in which response data are fully observed only if exceeding a certain threshold, often set to 0. This implies that $y_i = z_i\mathbb{1}(z_i > 0)$, with $(z_i \mid \boldsymbol{\beta}) \sim \text{N}(\mathbf{x}_i^\intercal\boldsymbol{\beta}, \sigma^2)$, independently for $i = 1, \ldots, n$. Such a formulation yields the joint likelihood

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\beta}) &\propto \prod_{i=1}^{n} \phi(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}; \sigma^2)^{\mathbb{1}(y_i>0)}\Phi(-\mathbf{x}_i^\intercal\boldsymbol{\beta}; \sigma^2)^{\mathbb{1}(y_i=0)}, \\ &= \phi_{n_1}(\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}; \sigma^2\mathbf{I}_{n_1})\Phi_{n_0}(-\mathbf{X}_0\boldsymbol{\beta}; \sigma^2\mathbf{I}_{n_0}), \end{aligned} \tag{1.9}$$

where n_1 and n_0 denote the number of fully observed and censored units, respectively, whereas $\mathbf{y}_1$, $\mathbf{X}_1$ and $\mathbf{X}_0$ are the response vectors and design matrices associated with these two subsets of units. This likelihood can be again expressed as a special example of equation (1.1) by letting $\bar{n}_1 = n_1$, $\bar{n}_0 = n_0$, $\bar{\mathbf{y}}_1 = \mathbf{y}_1$, $\bar{\mathbf{y}}_0 = \mathbf{0}$, $\bar{\mathbf{X}}_1 = \mathbf{X}_1$, $\bar{\mathbf{X}}_0 = -\mathbf{X}_0$, $\bar{\boldsymbol{\Sigma}}_1 = \sigma^2\mathbf{I}_{n_1}$ and $\bar{\boldsymbol{\Sigma}}_0 = \sigma^2\mathbf{I}_{n_0}$.

The above result also holds for several subsequent extensions of the original tobit model (Tobin, 1958) which include more elaborated censoring mechanisms, possibly relying on

multivariate Gaussian utilities. Such generalizations, often known in the literature as type II, III, IV, and V tobit models, are carefully discussed in Amemiya (1984) and all induce likelihoods which can be written as the product of Gaussian densities and cumulative distribution functions evaluated at suitable linear combinations of the coefficients $\boldsymbol{\beta}$. This common structure allows again to readily express such extensions as special cases of the general form in (1.1). The inclusion of multivariate extensions is also straightforward under similar reasoning considered in (1.3) and (1.6).

As for the models presented in Sections 1.2.1–1.2.2, also tobit regression admits extensions to skewed contexts. This generalization has been explored, for example, by Hutton & Stanghellini (2011) who replace the Gaussian assumption $(z_i \mid \boldsymbol{\beta}) \sim \text{N}(\mathbf{x}_i^\intercal\boldsymbol{\beta}, \sigma^2)$ with $(z_i \mid \boldsymbol{\beta}) \sim \text{SN}(\mathbf{x}_i^\intercal\boldsymbol{\beta}, \sigma^2, \alpha)$, independently for $i = 1, \ldots, n$. Recalling the derivations for the skewed extensions of the models in Sections 1.2.1–1.2.2, this assumption implies that the contribution to the likelihood for the i–th unit is proportional to $(\phi(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}; \sigma^2)\Phi[\alpha(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}); \sigma^2])^{\mathbb{1}(y_i>0)}(\Phi_2[(-\mathbf{x}_i^\intercal\boldsymbol{\beta}, 0)^\intercal; \text{diag}(\sigma^2, 1) - \sigma\alpha(1+\alpha^2)^{-1/2}(\mathbf{1}_2\mathbf{1}_2^\intercal - \mathbf{I}_2)])^{\mathbb{1}(y_i=0)}$, where $\text{diag}(\sigma^2, 1) = \boldsymbol{\Sigma}$ and $\sigma\alpha(1+\alpha^2)^{-1/2}(\mathbf{1}_2\mathbf{1}_2^\intercal - \mathbf{I}_2) = \boldsymbol{\Lambda}$. Therefore

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\beta}) \propto \prod_{i=1}^n & \Big(\phi(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}; \sigma^2)\Phi[\alpha(y_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}); \sigma^2]\Big)^{\mathbb{1}(y_i>0)} \cdot \\ & \cdot \Big(\Phi_2[(-\mathbf{x}_i^\intercal\boldsymbol{\beta}, 0)^\intercal; \boldsymbol{\Sigma} - \boldsymbol{\Lambda}]\Big)^{\mathbb{1}(y_i=0)} \\ = \phi_{n_1}(\mathbf{y}_1 - & \mathbf{X}_1\boldsymbol{\beta}; \sigma^2\mathbf{I}_{n_1})\Phi_{n_1+2n_0}(\alpha[\mathbf{y}_1^\intercal, \mathbf{0}^\intercal]^\intercal - (\alpha\mathbf{X}_1^\intercal, \mathbf{X}_0^\intercal)^\intercal\boldsymbol{\beta}; \boldsymbol{\Sigma}_0), \end{aligned} \tag{1.10}$$

where n_1, n_0, $\mathbf{y}_1$ and $\mathbf{X}_1$ are defined as in (1.9), whereas $\mathbf{X}_0$ is a $2n_0 \times p$ design matrix obtained by stacking $2 \times p$ row blocks $\mathbf{X}_i = (\mathbf{x}_i, \mathbf{0})^\intercal$ for those units with $y_i = 0$, while $\boldsymbol{\Sigma}_0$ is a block–diagonal matrix with blocks $\boldsymbol{\Sigma}_{0[1,1]} = \sigma^2\mathbf{I}_{n_1}$, $\boldsymbol{\Sigma}_{0[2,2]} = \mathbf{I}_{n_0} \otimes \boldsymbol{\Sigma} - \mathbf{I}_{n_0} \otimes \boldsymbol{\Lambda}$. Hence, to express (1.10) as a particular case of (1.1) it suffices to set $\bar{n}_1 = n_1$, $\bar{n}_0 = n_1 + 2n_0$, $\bar{\mathbf{y}}_1 = \mathbf{y}_1$, $\bar{\mathbf{y}}_0 = \alpha[\mathbf{y}_1^\intercal, \mathbf{0}^\intercal]^\intercal$, $\bar{\mathbf{X}}_1 = \mathbf{X}_1$, $\bar{\mathbf{X}}_0 = -(\alpha\mathbf{X}_1^\intercal, \mathbf{X}_0^\intercal)^\intercal$, $\bar{\boldsymbol{\Sigma}}_1 = \sigma^2\mathbf{I}_{n_1}$ and $\bar{\boldsymbol{\Sigma}}_0 = \boldsymbol{\Sigma}_0$. Recalling discussions in Sections 1.2.1–1.2.2, these derivations can be directly applied to incorporate skewness in type II–V tobit models (Amemiya, 1984), also under more general distributions which extend the original skew–normal.

1.2.4 Further generalizations

Although the models discussed in Sections 1.2.1–1.2.3 cover the most widely–implemented formulations in the literature, as highlighted in Sections 1.2.4–1.2.4 several additional extensions of these representations to non–linear, dynamic and other contexts can be reframed within the likelihood in (1.1).

Inclusion of generic thresholds

All the results presented in Sections 1.2.1–1.2.3 hold, under minor changes, when replacing the commonly–used 0 threshold with a generic one z_{T}, possibly varying with units. For

instance, in probit regression this modification implies that $\text{pr}(y_i = 1 \mid \boldsymbol{\beta}) = \Phi(-z_{\text{T}} + \mathbf{x}_i^\intercal \boldsymbol{\beta})$, thus providing the joint likelihood $\prod_{i=1}^n \Phi[(2y_i - 1)(-z_{\text{T}} + \mathbf{x}_i^\intercal \boldsymbol{\beta})] = \Phi_n(-z_{\text{T}}(2\mathbf{y} - \mathbf{1}_n) + \text{diag}(2\mathbf{y} - \mathbf{1}_n)\mathbf{X}\boldsymbol{\beta}; \mathbf{I}_n)$, which coincides with expression (1.1) after letting $\bar{n}_1 = 0$, $\bar{n}_0 = n$, $\bar{\mathbf{y}}_0 = -z_{\text{T}}(2\mathbf{y} - \mathbf{1}_n)$, $\bar{\mathbf{X}}_0 = \text{diag}(2\mathbf{y} - \mathbf{1}_n)\mathbf{X}$ and $\bar{\boldsymbol{\Sigma}}_0 = \mathbf{I}_n$. Similar derivations apply to multivariate probit, multinomial probit, tobit, and their skewed extensions.

By contrast, models relying on truncations to a finite interval of the form $[z_{1\text{T}}, z_{2\text{T}}]$ do not induce likelihoods which can be rewritten as in (1.1). Nonetheless, these versions are less frequent than those presented in Sections 1.2.1–1.2.3 and, as discussed in Section 1.6, the SUN conjugacy results presented for the general class of models whose likelihoods admit representation (1.1), are useful to motivate similar extensions for a generic truncation mechanism. In fact, as discussed in Arellano-Valle et al. (2006), the SUN family belongs itself to an even more general class of selection distributions (SLCT) whose construction rely on cumulative distribution functions evaluated at generic intervals. This result has been recently leveraged by Kowal (2021) and King & Kowal (2021) to extend the original SUN conjugacy properties presented by Durante (2019) and Fasano et al. (2021) for probit regression and its multivariate dynamic extensions, respectively, to rounded/categorical data situations where truncation is in finite intervals (e.g., Jeliazkov et al., 2008). These modifications can be extended to prove the SLCT conjugacy for generalizations of (1.1) which admit truncation to any finite interval.

Inclusion of non–linear effects

Another key extension of the models presented in Sections 1.2.1–1.2.3 can be obtained by including non–linearities in the predictor. A common solution to accomplish this goal is to replace the linear predictor $f(\mathbf{x}_i) = \mathbf{x}_i^\intercal \boldsymbol{\beta}$ with a generic basis expansion $f(\mathbf{x}_i) = \mathbf{g}(\mathbf{x}_i)^\intercal \boldsymbol{\beta}$, where $\mathbf{g}(\mathbf{x}_i) = [g_1(\mathbf{x}_i), \ldots, g_k(\mathbf{x}_i)]^\intercal$ are pre–specified non–linear basis functions, such as splines (e.g., Holmes & Mallick, 2001; Lang & Brezger, 2004). Including this extension within the general framework in (1.1) poses no difficulties since it suffices to replicate the derivations for the models presented in Sections 1.2.1–1.2.3 with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$ replaced by $\tilde{\mathbf{x}}_i = [g_1(\mathbf{x}_i), \ldots, g_k(\mathbf{x}_i)]^\intercal$, for each $i = 1, \ldots, n$.

Alternatively, it is possible to model directly $[f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\intercal$ through Gaussian processes (e.g., Rasmussen & Williams, 2006). This direction has been often explored in the context of the models presented in Sections 1.2.1–1.2.3 (e.g., Kuss et al., 2005; De Oliveira, 2005; Girolami & Rogers, 2006; Nickisch & Rasmussen, 2008; Riihimäki et al., 2014; Cao et al., 2022; Benavoli et al., 2020, 2021), and can be also reframed within equation (1.1). In fact, assuming, without loss of generality, no overlap in $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the Gaussian process construction with mean function $m(\cdot)$ and covariance kernel $K(\cdot, \cdot)$ implies that the vector $[f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\intercal$ is jointly distributed as a $\text{N}_n(\boldsymbol{\xi}, \boldsymbol{\Omega})$ with $\boldsymbol{\xi} = [m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n)]^\intercal$ and covariance matrix $\boldsymbol{\Omega}$ having generic entries $\boldsymbol{\Omega}_{ii'} = K(\mathbf{x}_i, \mathbf{x}_{i'})$, for each $i = 1, \ldots, n$ and $i' = 1, \ldots, n$. This representation can be alternatively rewritten as $\tilde{\mathbf{X}}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\intercal \sim \text{N}_n(\boldsymbol{\xi}, \boldsymbol{\Omega})$ and $\tilde{\mathbf{X}} = \mathbf{I}_n$. Therefore, letting $\tilde{\mathbf{x}}_i$ denote an $n \times 1$ vector with

value 1 in position i and 0 elsewhere, for each $i = 1, \dots, n$, it is possible to consider Gaussian process extensions of the models in Sections 1.2.1–1.2.3 while still remaining within the general framework in (1.1).
The above results can be naturally adapted to multivariate settings, such as those in (1.3) and (1.6).

Inclusion of dynamic structure

Dynamic versions of the models in Sections 1.2.1–1.2.3 are common in the literature (e.g., Manrique & Shephard, 1998; Andrieu & Doucet, 2002; Naveau et al., 2005; Chib & Jeliazkov, 2006; Soyer & Sung, 2013; Fasano et al., 2021). These extensions often appear as generalizations of the original dynamic linear model having observation equation $(\mathbf{y}_t \mid \boldsymbol{\beta}_t) \sim \mathrm{N}_m(\mathbf{X}_t\boldsymbol{\beta}_t, \boldsymbol{\Sigma}_t)$, independently for each time $t = 1, \dots, n$, and state equation $(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}) \sim \mathrm{N}_p(\mathbf{G}_t\boldsymbol{\beta}_{t-1}, \mathbf{W}_t)$, independently for $t = 1, \dots, n$, where $\mathbf{X}_t$, $\boldsymbol{\Sigma}_t$, $\mathbf{G}_t$, and $\mathbf{W}_t$ are known system matrices, whereas $\boldsymbol{\beta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0)$. This building–block construction implies that the contribution to the likelihood of $\mathbf{y}_t$, for every time $t = 1, \dots, n$, is $p(\mathbf{y}_t \mid \boldsymbol{\beta}) = \phi_m(\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta}_t; \boldsymbol{\Sigma}_t) = \phi_m(\mathbf{y}_t - \tilde{\mathbf{X}}_t\boldsymbol{\beta}; \boldsymbol{\Sigma}_t)$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\intercal, \dots, \boldsymbol{\beta}_n^\intercal)^\intercal$ and $\tilde{\mathbf{X}}_t = \mathbf{v}_t^\intercal \otimes \mathbf{X}_t$, with $\mathbf{v}_t$ denoting a $n \times 1$ indicator vector having value 1 in position t and 0 elsewhere. Therefore, such a representation can be directly interpreted as a particular version of the multivariate linear regression model in (1.3) with covariance matrix possibly changing across the time units. Such a connection allows to directly recast the joint likelihood $p(\mathbf{y} \mid \boldsymbol{\beta})$ of $\mathbf{y} = (\mathbf{y}_1^\intercal, \dots, \mathbf{y}_n^\intercal)^\intercal$ within (1.1). Clearly, this result holds for any subsequence $\mathbf{y}_{1:t}^\intercal = (\mathbf{y}_1^\intercal, \dots, \mathbf{y}_t^\intercal)^\intercal$, with $t = 1, \dots, n$, thereby facilitating online derivation of filtering $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$, predictive $p(\boldsymbol{\beta}_{t+1} \mid \mathbf{y}_{1:t})$ and smoothing $p(\boldsymbol{\beta} \mid \mathbf{y})$ distributions via the Gaussian–Gaussian conjugacy implied by the observation and state equations (Kalman, 1960).

The above results have been recently extended by Fasano et al. (2021) to derive the first analog of the classical Kalman filter (Kalman, 1960) in the context of multivariate dynamic probit models with Gaussian states, leveraging the SUN–probit conjugacy properties proved in Durante (2019). Recalling Fasano et al. (2021) and adapting the notation to the one in this Chapter, the contribution to the likelihood of $\mathbf{y}_t$, for every time $t = 1, \dots, n$, can be expressed as $p(\mathbf{y}_t \mid \boldsymbol{\beta}) = \Phi_m(\mathbf{B}_t\mathbf{X}_t\boldsymbol{\beta}_t; \mathbf{B}_t\boldsymbol{\Sigma}_t\mathbf{B}_t)$, where $\mathbf{X}_t$ and $\mathbf{B}_t$ are defined as in (1.6), with i replaced by t, whereas $\boldsymbol{\Sigma}_t$ is a possibly time–varying covariance matrix among the latent utilities $(z_{t1}, \dots, z_{tm})^\intercal$. Recalling the derivations considered for the Gaussian dynamic setting, the expression for $p(\mathbf{y}_t \mid \boldsymbol{\beta})$ can be rewritten as $p(\mathbf{y}_t|\boldsymbol{\beta}) = \Phi_m(\tilde{\mathbf{X}}_t\boldsymbol{\beta}; \mathbf{B}_t\boldsymbol{\Sigma}_t\mathbf{B}_t)$, with $\tilde{\mathbf{X}}_t = \mathbf{v}_t^\intercal \otimes (\mathbf{B}_t\mathbf{X}_t)$, which shows again the direct connection between this dynamic formulation and its static counterpart in (1.6), thereby allowing to recast the induced joint likelihood for $\mathbf{y} = (\mathbf{y}_1^\intercal, \dots, \mathbf{y}_n^\intercal)^\intercal$ and its subsequences $\mathbf{y}_{1:t}^\intercal = (\mathbf{y}_1^\intercal, \dots, \mathbf{y}_t^\intercal)^\intercal$, $t = 1, \dots, n$, within expression (1.1).

These results clearly hold also for the dynamic extensions of models (1.2) and (1.5), which represent the univariate versions of (1.3) and (1.6), respectively, thus simply requir-

ing to set $m = 1$ in the above derivations. Similarly, multinomial probit (1.7) and tobit (1.9) observation equations, along with skewed extensions ((1.4), (1.8), (1.10)), can be again reframed within (1.1) since all these constructions are characterized by contributions to the likelihood for each time $t = 1, \dots, n$ having the same form of those associated with statistical units $i = 1, \dots, n$ in the static counterparts of such models presented in Sections 1.2.1–1.2.3, under suitable specifications of the design and covariance matrices.

As a last remark, it is worth emphasizing that equation (1.1) naturally encompasses any combination of the models discussed in Sections 1.2.1–1.2.4. For example, if $\mathbf{y}_i$ is defined as $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i2}, y_{i4})^\intercal$, where y_{i1}, y_{i2}, y_{i2} and y_{i4} are from models in (1.2), (1.5), (1.7) and (1.9), respectively, for each $i = 1, \dots, n$, then, leveraging the derivations in Sections 1.2.1–1.2.3, it directly follows that the joint likelihood for the vector $\mathbf{y} = (\mathbf{y}_1^\intercal, \dots, \mathbf{y}_n^\intercal)^\intercal$ still belongs to (1.1).

1.3 Conjugacy via unified skew–normal distributions

Sections 1.3.1–1.3.2 unify Bayesian inference for the whole family of models presented in Section 1.2 by proving that the general likelihood in (1.1) admits as conjugate priors the whole class of unified skew–normal distributions. Crucially, these variables include as special cases the commonly–assumed multivariate Gaussian priors for $\boldsymbol{\beta}$ in models (1.2)–(1.10), while extending such distributions in several directions. In this way, such a review not only unifies and extends a broad class of models within a single likelihood, but also enlarges the class of prior distributions which admit closed–form posteriors that facilitate Bayesian inference.

1.3.1 Unified skew–normal prior

Routine Bayesian implementations of the models in Sections 1.2.1–1.2.4 often assume multivariate Gaussian priors for $\boldsymbol{\beta}$, which are natural choices in Bayesian regression and, under the models presented in Sections 1.2.1–1.2.4, are further motivated by the Gaussian form of the latent utilities (e.g., Chib, 1992; Albert & Chib, 1993; McCulloch & Rossi, 1994; Holmes & Held, 2006; Girolami & Rogers, 2006; McCulloch & Rossi, 1994; Nobile, 1998; Chib & Greenberg, 1998; McCulloch et al., 2000; Albert & Chib, 2001; Imai & Van Dyk, 2005; Kuss et al., 2005; Riihimäki et al., 2014; Soyer & Sung, 2013; Chopin & Ridgway, 2017). Interestingly, these Gaussian priors are special cases of more general distributions which include asymmetric shapes in multivariate Gaussians by modifying the density of such variables through a skewness–inducing mechanism driven by the cumulative distribution function of another Gaussian. Key examples include multivariate skew–normals (Azzalini & Dalla Valle, 1996; Azzalini & Capitanio, 1999), extended multivariate skew–normals (Arnold & Beaver, 2000; Arnold et al., 2002) and closed skew–normals (González-Farias et al., 2004; Gupta et al., 2004), which have all been subsequently unified by Arellano-Valle

& Azzalini (2006) within a single general class, namely the unified skew–normal distribution. Accordingly, the vector $\boldsymbol{\beta}$ has $\text{SUN}_{\bar{p},\bar{n}}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ prior if its density is equal to

$$p(\boldsymbol{\beta}) = \phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \frac{\Phi_{\bar{n}}(\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta})}{\Phi_{\bar{n}}(\boldsymbol{\gamma}; \boldsymbol{\Gamma})}, \tag{1.11}$$

where $\bar{\boldsymbol{\Omega}}$ denotes the $\bar{p} \times \bar{p}$ correlation matrix associated with the covariance matrix $\boldsymbol{\Omega}$ which, in turn, can be expressed as $\boldsymbol{\Omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}$, with $\boldsymbol{\omega} = (\boldsymbol{\Omega} \odot \mathbf{I}_{\bar{p}})^{1/2}$. According to (1.11), skewness is induced in $\phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega})$ by multiplying such a density with the cumulative distribution function of a $\text{N}_{\bar{n}}(\mathbf{0}, \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta})$, evaluated at $\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi})$, whereas $\Phi_{\bar{n}}(\boldsymbol{\gamma}; \boldsymbol{\Gamma})$ corresponds to the normalizing constant. Note that when all the entries in the $\bar{p} \times \bar{n}$ skewness matrix $\boldsymbol{\Delta}$ are 0, the numerator in (1.11) reduces to $\Phi_{\bar{n}}(\boldsymbol{\gamma}; \boldsymbol{\Gamma})$, thereby allowing to obtain the classical Gaussian prior density $\phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega})$ as a special case of (1.11). The quantities $\bar{p}$ and $\bar{n}$ denote instead the dimensions of the density and the cumulative distribution function, respectively. Within the general class of formulations discussed in Sections 1.2.1–1.2.4, $\bar{p}$ refers to the dimension of $\boldsymbol{\beta}$ and, hence, can vary depending on the model considered. While in most cases $\bar{p}$ is equal to the number of predictors p, under specific constructions the two dimensions might differ. For instance, in the multinomial probit model presented in equation (1.7), $\bar{p}$ coincides with $p \cdot (L - 1)$.

Recalling Arellano-Valle & Azzalini (2006), the above SUN distribution also admits a generative construction that further clarifies the role of the parameters $\boldsymbol{\xi}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Delta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\Gamma}$, and provides key intuitions on the conjugacy properties of SUN priors under likelihood (1.1). In particular, let $\tilde{\boldsymbol{\beta}} \in \Re^{\bar{p}}$ and $\tilde{\mathbf{z}} \in \Re^{\bar{n}}$ denote two vectors jointly distributed as a unified skew-normal $\text{N}_{\bar{p}+\bar{n}}(\mathbf{0}, \boldsymbol{\Omega}^*)$, where $\boldsymbol{\Omega}^*$ is a $(\bar{p} + \bar{n}) \times (\bar{p} + \bar{n})$ correlation matrix with blocks $\boldsymbol{\Omega}^*_{[11]} = \bar{\boldsymbol{\Omega}}$, $\boldsymbol{\Omega}^*_{[22]} = \boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}^*_{[21]} = \boldsymbol{\Omega}^{*\intercal}_{[12]} = \boldsymbol{\Delta}^\intercal$, then $\bar{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}} \mid \tilde{\mathbf{z}} + \boldsymbol{\gamma} > \mathbf{0})$ is distributed as a $\text{SUN}_{\bar{p},\bar{n}}(\mathbf{0}, \bar{\boldsymbol{\Omega}}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, whereas $\boldsymbol{\beta} = \boldsymbol{\xi} + \boldsymbol{\omega}\bar{\boldsymbol{\beta}} \sim \text{SUN}_{\bar{p},\bar{n}}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ with density as in (1.11). Consistent with this generative representation, the parameters $\boldsymbol{\xi}$ and $\boldsymbol{\omega}$ control the location and the scale of the prior, whereas $\bar{\boldsymbol{\Omega}}$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Delta}$ regulate the dependence within $\tilde{\boldsymbol{\beta}}$, $\tilde{\mathbf{z}}$ and between these two random vectors, respectively. The term $\boldsymbol{\gamma}$ denotes instead the truncation threshold in the conditioning mechanism. Besides clarifying the role of the prior parameters, this representation also provides intuitions on the SUN conjugacy properties formalized in Section 1.3.2. In fact, according to such a construction, SUNs arise as conditional distributions in a generative mechanism that relies on partially–observed Gaussian latent variables $\tilde{\mathbf{z}}$. This formulation has direct connections with the posterior distribution for the $\boldsymbol{\beta}$ coefficients under the broad class of models presented in Sections 1.2.1–1.2.4 that is also defined, through Bayes rule, via a conditioning operation relying on a set of partially or fully observed Gaussian latent utilities.

As clarified in Section 1.3.2, these conjugacy properties are also beneficial for posterior inference. Recalling Arellano-Valle & Azzalini (2006); Azzalini & Bacchieri (2010); Gupta et al. (2013); Azzalini & Capitanio (2013), and Arellano-Valle & Azzalini (2021) SUNs share a number of common properties with multivariate Gaussians. These include closure un-

der marginalization, linear combinations and conditioning, along with the availability of closed–form expressions for the moment generating function and additive representations via linear combinations of multivariate Gaussians and truncated normals. Due to SUN conjugacy shown in Theorem 1.2, all these properties can facilitate point estimation, uncertainty quantification, model selection and prediction under the SUN posterior associated with the general likelihood in (1.1) which encompasses the models in Sections 1.2.1–1.2.4, thus providing important advancements for a broad class of models, under a similarly wide family of priors that extends multivariate Gaussians. Key examples of priors of potential interest which belong to the SUN family are univariate skew–normals (Azzalini, 1985) for each coefficient β_j, $j = 1, \dots, \bar{p}$, or multivariate skew–normals (Azzalini & Dalla Valle, 1996; Azzalini & Capitanio, 1999), extended multivariate skew–normals (Arnold & Beaver, 2000; Arnold et al., 2002) and closed skew–normals (González-Farias et al., 2004; Gupta et al., 2004) for the joint vector $\boldsymbol{\beta}$.

1.3.2 Unified skew–normal posterior and its properties

Theorem 1.2 unifies and extends recent model–specific conjugacy findings by proving SUN conjugacy for any statistical model whose likelihood can be expressed as in equation (1.1). The proof of Theorem 1.2 combines original results on SUN conjugacy in probit models (Durante, 2019, Corollary 4) with the following Lemma, which shows that SUN priors are also conjugate to Gaussian linear regression.

Lemma 1.1. *Let* $p(\bar{\mathbf{y}}_1 \mid \boldsymbol{\beta}) = \phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_1)$ *and assume that* $\boldsymbol{\beta}$ *is assigned a prior distribution* $\text{SUN}_{\bar{p},\bar{n}}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ *with density* $p(\boldsymbol{\beta})$ *as in* (1.11). *Then, we have that* $(\boldsymbol{\beta} \mid \bar{\mathbf{y}}_1) \sim \text{SUN}_{\bar{p},\bar{n}}(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\Delta}_1, \boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_1)$, *with*

$$\boldsymbol{\xi}_1 = (\boldsymbol{\Omega}^{-1} + \bar{\mathbf{X}}_1^\intercal \bar{\boldsymbol{\Sigma}}_1^{-1} \bar{\mathbf{X}}_1)^{-1}(\boldsymbol{\Omega}^{-1}\boldsymbol{\xi} + \bar{\mathbf{X}}_1^\intercal \bar{\boldsymbol{\Sigma}}_1^{-1} \bar{\mathbf{y}}_1), \qquad \boldsymbol{\Omega}_1 = (\boldsymbol{\Omega}^{-1} + \bar{\mathbf{X}}_1^\intercal \bar{\boldsymbol{\Sigma}}_1^{-1} \bar{\mathbf{X}}_1)^{-1},$$
$$\boldsymbol{\gamma}_1 = \mathbf{s}_1^{-1}[\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\xi}_1 - \boldsymbol{\xi})], \qquad \boldsymbol{\Delta}_1 = \bar{\boldsymbol{\Omega}}_1 \boldsymbol{\omega}_1 \boldsymbol{\omega}^{-1} \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta} \mathbf{s}_1^{-1},$$
$$\boldsymbol{\Gamma}_1 = \mathbf{s}_1^{-1}[\boldsymbol{\Gamma} + \boldsymbol{\Delta}^\intercal (\bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1} \boldsymbol{\Omega}_1 \boldsymbol{\omega}^{-1} \bar{\boldsymbol{\Omega}}^{-1} - \bar{\boldsymbol{\Omega}}^{-1})\boldsymbol{\Delta}]\mathbf{s}_1^{-1},$$

where $\mathbf{s}_1 = ([\boldsymbol{\Gamma} + \boldsymbol{\Delta}^\intercal (\bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1} \boldsymbol{\Omega}_1 \boldsymbol{\omega}^{-1} \bar{\boldsymbol{\Omega}}^{-1} - \bar{\boldsymbol{\Omega}}^{-1})\boldsymbol{\Delta}] \odot \mathbf{I}_{\bar{n}})^{1/2}$.

Note that in Lemma 1.1 the rescaling operated by $\mathbf{s}_1$ is required to ensure that the matrix $\boldsymbol{\Omega}_1^*$ with blocks $\boldsymbol{\Omega}_{1[11]}^* = \bar{\boldsymbol{\Omega}}_1$, $\boldsymbol{\Omega}_{1[22]}^* = \boldsymbol{\Gamma}_1$ and $\boldsymbol{\Omega}_{1[21]}^* = \boldsymbol{\Omega}_{1[12]}^{*\intercal} = \boldsymbol{\Delta}_1^\intercal$ is a correlation matrix, as in the original formulation by Arellano-Valle & Azzalini (2006). Although this constraint is useful to avoid identifiability issues in frequentist settings, such problems are less of a concern in the Bayesian setting considered in this Chapter, since the parameters of the SUN posterior are function of the observed data and of the known prior hyperparameters. Nonetheless, maintaining this constraint is still useful to inherit results of the original SUN and to avoid identifiability issues in prior elicitation.

Proof. To prove Lemma 1.1, first notice that, by Bayes rule, $p(\boldsymbol{\beta} \mid \bar{\mathbf{y}}_1) \propto p(\boldsymbol{\beta}) p(\bar{\mathbf{y}}_1 \mid \boldsymbol{\beta})$, where $p(\bar{\mathbf{y}}_1 \mid \boldsymbol{\beta}) = \phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_1)$, whereas $p(\boldsymbol{\beta})$ is the SUN density in (1.11). Leveraging

Gaussian–Gaussian conjugacy, it follows that the product between $\phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_1)$ and the density term $\phi_{\bar{p}}(\boldsymbol{\beta}-\boldsymbol{\xi}; \boldsymbol{\Omega})$ in (1.11) is proportional to $\phi_{\bar{p}}[\boldsymbol{\beta}-(\boldsymbol{\Omega}^{-1}+\bar{\mathbf{X}}_1^\intercal\bar{\boldsymbol{\Sigma}}_1^{-1}\bar{\mathbf{X}}_1)^{-1}(\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}+\bar{\mathbf{X}}_1^\intercal\bar{\boldsymbol{\Sigma}}_1^{-1}\bar{\mathbf{y}}_1); (\boldsymbol{\Omega}^{-1}+\bar{\mathbf{X}}_1^\intercal\bar{\boldsymbol{\Sigma}}_1^{-1}\bar{\mathbf{X}}_1)^{-1}] = \phi_{\bar{p}}(\boldsymbol{\beta}-\boldsymbol{\xi}_1; \boldsymbol{\Omega}_1)$, where $\boldsymbol{\Omega}_1 = (\boldsymbol{\Omega}^{-1}+\bar{\mathbf{X}}_1^\intercal\bar{\boldsymbol{\Sigma}}_1^{-1}\bar{\mathbf{X}}_1)^{-1} = \boldsymbol{\omega}_1\bar{\boldsymbol{\Omega}}_1\boldsymbol{\omega}_1$ and $\boldsymbol{\xi}_1 = (\boldsymbol{\Omega}^{-1}+\bar{\mathbf{X}}_1^\intercal\bar{\boldsymbol{\Sigma}}_1^{-1}\bar{\mathbf{X}}_1)^{-1}(\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}+\bar{\mathbf{X}}_1^\intercal\bar{\boldsymbol{\Sigma}}_1^{-1}\bar{\mathbf{y}}_1)$. Therefore, $p(\boldsymbol{\beta} \mid \bar{\mathbf{y}}_1)$ is proportional to the product between this updated Gaussian density and the cumulative distribution function term $\Phi_{\bar{n}}(\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta})$ of the SUN density in (1.11), which can be also re–expressed as $\Phi_{\bar{n}}[\mathbf{s}_1^{-1}\boldsymbol{\gamma}+\mathbf{s}_1^{-1}\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}); \mathbf{s}_1^{-1}(\boldsymbol{\Gamma}-\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta})\mathbf{s}_1^{-1}]$, where $\mathbf{s}_1^{-1}$ is defined as in Lemma 1.1. To prove that this product yields the SUN kernel in Lemma 1.1, rewrite $\mathbf{s}_1^{-1}\boldsymbol{\gamma} + \mathbf{s}_1^{-1}\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi})$ as $\mathbf{s}_1^{-1}\boldsymbol{\gamma} - \mathbf{s}_1^{-1}\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}\boldsymbol{\xi} + \mathbf{s}_1^{-1}\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}\boldsymbol{\beta}$, and then sum and subtract $\mathbf{s}_1^{-1}\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}\boldsymbol{\xi}_1$ inside this expression to obtain $\mathbf{s}_1^{-1}[\boldsymbol{\gamma}+\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\xi}_1-\boldsymbol{\xi})]+\mathbf{s}_1^{-1}\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_1) = \boldsymbol{\gamma}_1+\mathbf{s}_1^{-1}\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_1)$. Let us now replace in this formula the term $\mathbf{s}_1^{-1}\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}$ with $\mathbf{s}_1^{-1}\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}\boldsymbol{\omega}_1\bar{\boldsymbol{\Omega}}_1\bar{\boldsymbol{\Omega}}_1^{-1}\boldsymbol{\omega}_1^{-1} = \mathbf{s}_1^{-1}\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1} = \boldsymbol{\Delta}_1^\intercal\bar{\boldsymbol{\Omega}}_1{}^{-1}\boldsymbol{\omega}_1^{-1}$, where $\boldsymbol{\Delta}_1 = \bar{\boldsymbol{\Omega}}_1\boldsymbol{\omega}_1\boldsymbol{\omega}^{-1}\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta}\mathbf{s}_1^{-1}$. To conclude the proof, note that the correlation matrix within the cumulative distribution function term can be also rewritten as $\mathbf{s}_1^{-1}(\boldsymbol{\Gamma}-\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta})\mathbf{s}_1^{-1} = \mathbf{s}_1^{-1}(\boldsymbol{\Gamma}-\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta})\mathbf{s}_1^{-1}+\boldsymbol{\Delta}_1^\intercal\bar{\boldsymbol{\Omega}}_1^{-1}\boldsymbol{\Delta}_1-\boldsymbol{\Delta}_1^\intercal\bar{\boldsymbol{\Omega}}_1^{-1}\boldsymbol{\Delta}_1 = \mathbf{s}_1^{-1}[\boldsymbol{\Gamma} + \boldsymbol{\Delta}^\intercal(\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}\boldsymbol{\Omega}_1\boldsymbol{\omega}^{-1}\bar{\boldsymbol{\Omega}}^{-1} - \bar{\boldsymbol{\Omega}}^{-1})\boldsymbol{\Delta}]\mathbf{s}_1^{-1} - \boldsymbol{\Delta}_1^\intercal\bar{\boldsymbol{\Omega}}_1^{-1}\boldsymbol{\Delta}_1$, which corresponds to $\boldsymbol{\Gamma}_1 - \boldsymbol{\Delta}_1^\intercal\bar{\boldsymbol{\Omega}}_1^{-1}\boldsymbol{\Delta}_1$, with $\boldsymbol{\Gamma}_1$ defined as in Lemma 1.1. This proves that the kernel $p(\boldsymbol{\beta})p(\bar{\mathbf{y}}_1 \mid \boldsymbol{\beta})$ of the posterior coincides with that of a SUN having parameters $\boldsymbol{\xi}_1$, $\boldsymbol{\Omega}_1$, $\boldsymbol{\Delta}_1$, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\Gamma}_1$ specified as in Lemma 1.1. □

Leveraging Lemma 1.1 and adapting Corollary 4 in Durante (2019), it is now possible to prove the general SUN conjugacy result stated in Theorem 1.2.

Theorem 1.2. *Let* $p(\mathbf{y} \mid \boldsymbol{\beta}) = p(\bar{\mathbf{y}}_1 \mid \boldsymbol{\beta})p(\bar{\mathbf{y}}_0 \mid \boldsymbol{\beta}) \propto \phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_1)\Phi_{\bar{n}_0}(\bar{\mathbf{y}}_0 + \bar{\mathbf{X}}_0\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_0)$ *as in* (1.1)*, and assume that* $\boldsymbol{\beta}$ *is assigned a* $\text{SUN}_{\bar{p},\bar{n}}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ *prior distribution with density* $p(\boldsymbol{\beta})$ *as in* (1.11)*. Then,* $(\boldsymbol{\beta} \mid \mathbf{y}) \sim \text{SUN}_{\bar{p},\bar{n}+\bar{n}_0}(\boldsymbol{\xi}_{\text{post}}, \boldsymbol{\Omega}_{\text{post}}, \boldsymbol{\Delta}_{\text{post}}, \boldsymbol{\gamma}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}})$*, with posterior parameters*

$$\boldsymbol{\xi}_{\text{post}} = \boldsymbol{\xi}_1, \quad \boldsymbol{\Omega}_{\text{post}} = \boldsymbol{\Omega}_1, \quad \boldsymbol{\Delta}_{\text{post}} = (\boldsymbol{\Delta}_1, \bar{\boldsymbol{\Omega}}_1\boldsymbol{\omega}_1\bar{\mathbf{X}}_0^\intercal\mathbf{s}_0^{-1}), \quad \boldsymbol{\gamma}_{\text{post}} = [\boldsymbol{\gamma}_1^\intercal, (\bar{\mathbf{y}}_0 + \bar{\mathbf{X}}_0\boldsymbol{\xi}_{\text{post}})^\intercal\mathbf{s}_0^{-1}]^\intercal,$$

and $\boldsymbol{\Gamma}_{\text{post}}$ *characterizing a full–rank* $(\bar{n} + \bar{n}_0) \times (\bar{n} + \bar{n}_0)$ *correlation matrix with blocks* $\boldsymbol{\Gamma}_{\text{post}[11]} = \boldsymbol{\Gamma}_1$, $\boldsymbol{\Gamma}_{\text{post}[22]} = \mathbf{s}_0^{-1}(\bar{\mathbf{X}}_0\boldsymbol{\Omega}_1\bar{\mathbf{X}}_0^\intercal + \bar{\boldsymbol{\Sigma}}_0)\mathbf{s}_0^{-1}$*, and* $\boldsymbol{\Gamma}_{\text{post}[21]} = \boldsymbol{\Gamma}_{\text{post}[12]}^\intercal = \mathbf{s}_0^{-1}\bar{\mathbf{X}}_0\boldsymbol{\omega}_1\boldsymbol{\Delta}_1$*, where* $\mathbf{s}_0 = ([\bar{\mathbf{X}}_0\boldsymbol{\Omega}_1\bar{\mathbf{X}}_0^\intercal + \bar{\boldsymbol{\Sigma}}_0] \odot \mathbf{I}_{\bar{n}_0})^{1/2}$*, while* $\boldsymbol{\xi}_1$, $\boldsymbol{\Omega}_1$, $\boldsymbol{\Delta}_1$, $\boldsymbol{\gamma}_1$ *and* $\boldsymbol{\Gamma}_1$ *are defined as in Lemma 1.1, namely*

$$\begin{aligned}
&\boldsymbol{\xi}_1 = (\boldsymbol{\Omega}^{-1}+\bar{\mathbf{X}}_1^\intercal\bar{\boldsymbol{\Sigma}}_1^{-1}\bar{\mathbf{X}}_1)^{-1}(\boldsymbol{\Omega}^{-1}\boldsymbol{\xi} + \bar{\mathbf{X}}_1^\intercal\bar{\boldsymbol{\Sigma}}_1^{-1}\bar{\mathbf{y}}_1), && \boldsymbol{\Omega}_1 = (\boldsymbol{\Omega}^{-1}+\bar{\mathbf{X}}_1^\intercal\bar{\boldsymbol{\Sigma}}_1^{-1}\bar{\mathbf{X}}_1)^{-1},\\
&\boldsymbol{\gamma}_1 = \mathbf{s}_1^{-1}[\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\xi}_1 - \boldsymbol{\xi})], && \boldsymbol{\Delta}_1 = \bar{\boldsymbol{\Omega}}_1\boldsymbol{\omega}_1\boldsymbol{\omega}^{-1}\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta}\mathbf{s}_1^{-1},\\
&\boldsymbol{\Gamma}_1 = \mathbf{s}_1^{-1}[\boldsymbol{\Gamma} + \boldsymbol{\Delta}^\intercal(\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}\boldsymbol{\Omega}_1\boldsymbol{\omega}^{-1}\bar{\boldsymbol{\Omega}}^{-1} - \bar{\boldsymbol{\Omega}}^{-1})\boldsymbol{\Delta}]\mathbf{s}_1^{-1}, &&
\end{aligned}$$

with $\mathbf{s}_1 = ([\boldsymbol{\Gamma} + \boldsymbol{\Delta}^\intercal(\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}\boldsymbol{\Omega}_1\boldsymbol{\omega}^{-1}\bar{\boldsymbol{\Omega}}^{-1} - \bar{\boldsymbol{\Omega}}^{-1})\boldsymbol{\Delta}] \odot \mathbf{I}_{\bar{n}})^{1/2}$.

Theorem 1.2 crucially encompasses all available conjugacy results for SUN distributions under specific models within the broader family analyzed, while extending such findings to other key formulations. For example, setting $\bar{p} = p$, $\bar{n}_1 = 0$, $\bar{n}_0 = n$, $\bar{\mathbf{y}}_0 = \mathbf{0}$, $\bar{\mathbf{X}}_0 = \text{diag}(2\mathbf{y}-\mathbf{1}_n)\mathbf{X}$ and $\bar{\boldsymbol{\Sigma}}_0 = \mathbf{I}_n$ as in model (1.5), and substituting these quantities within the

expressions in Theorem 1.2, would yield a $\text{SUN}_{p,\bar{n}+n}(\boldsymbol{\xi}_{\text{post}}, \boldsymbol{\Omega}_{\text{post}}, \boldsymbol{\Delta}_{\text{post}}, \boldsymbol{\gamma}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}})$ posterior with parameters as in Corollary 4 by Durante (2019). Theorem 1 in Durante (2019) is instead recovered under the additional constraint $\bar{n} = 0$, which implies a Gaussian prior. Note that when $\bar{n}_1 = 0$ the associated quantities $\bar{\mathbf{y}}_1$, $\bar{\mathbf{X}}_1$ and $\bar{\boldsymbol{\Sigma}}_1$ are not defined and simply need to be removed from the formulas in Theorem 1.2. The same reasoning holds for $\bar{\mathbf{y}}_0$, $\bar{\mathbf{X}}_0$ and $\bar{\boldsymbol{\Sigma}}_0$ when $\bar{n}_0 = 0$, and for $\boldsymbol{\Delta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\Omega}$ if $\bar{n} = 0$. For instance, setting $\bar{n}_0 = 0$ in Theorem 1.2 yields directly to Lemma 1.1. Similarly, the SUN conjugacy results for multinomial probit (Fasano & Durante, 2022), dynamic multivariate probit (Fasano et al., 2021), Gaussian processes (Cao et al., 2022), and skewed Gaussian processes under linear models, affine probit and combinations of these two formulations (Benavoli et al., 2020, 2021) can be readily obtained from Theorem 1.2 under the settings in Sections 1.2.1–1.2.4 for the quantities defining the likelihood in (1.1). Interestingly, also results outside the regression context, such as those proved by Canale et al. (2016) for multivariate skew–normal likelihoods with Gaussian or skew–normal priors on the shape parameter, can be recasted within Theorem 1.2. Besides encompassing already available findings, Theorem 1.2 provides novel conjugacy results also in previously–unexplored settings, such as in tobit regression and in models relying on skewed utilities.

Proof. The proof for Theorem 1.2 simply requires to combine Lemma 1.1 with an adaptation of Corollary 4 in Durante (2019). In particular, by direct application of the Bayes rule, it follows that $p(\boldsymbol{\beta} \mid \mathbf{y}) \propto p(\boldsymbol{\beta})p(\mathbf{y} \mid \boldsymbol{\beta}) \propto p(\boldsymbol{\beta})p(\bar{\mathbf{y}}_1 \mid \boldsymbol{\beta})p(\bar{\mathbf{y}}_0 \mid \boldsymbol{\beta})$. Hence, the posterior $p(\boldsymbol{\beta} \mid \mathbf{y}) = p(\boldsymbol{\beta} \mid \bar{\mathbf{y}}_1, \bar{\mathbf{y}}_0) \propto [p(\boldsymbol{\beta})p(\bar{\mathbf{y}}_1 \mid \boldsymbol{\beta})]p(\bar{\mathbf{y}}_0 \mid \boldsymbol{\beta})$ can be obtained by first updating the SUN prior $p(\boldsymbol{\beta})$ with $p(\bar{\mathbf{y}}_1 \mid \boldsymbol{\beta}) \propto \phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_1)$, and then use such a conditional density $p(\boldsymbol{\beta} \mid \bar{\mathbf{y}}_1)$ as an intermediate prior to be updated with the likelihood $p(\bar{\mathbf{y}}_0 \mid \boldsymbol{\beta}) \propto \Phi_{\bar{n}_0}(\bar{\mathbf{y}}_0 + \bar{\mathbf{X}}_0\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_0)$ of $\bar{\mathbf{y}}_0$ for obtaining the final posterior. By direct application of Lemma 1.1, it follows that $p(\boldsymbol{\beta} \mid \bar{\mathbf{y}}_1)$ is the density of the $\text{SUN}_{\bar{p},\bar{n}}(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\Delta}_1, \boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_1)$ with parameters defined as in Lemma 1.1. Therefore, to conclude the proof, it is sufficient to prove that the updating of this intermediate prior with the likelihood $p(\bar{\mathbf{y}}_0 \mid \boldsymbol{\beta}) \propto \Phi_{\bar{n}_0}(\bar{\mathbf{y}}_0 + \bar{\mathbf{X}}_0\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_0)$ for $\bar{\mathbf{y}}_0$ yields again to a $\text{SUN}_{\bar{p},\bar{n}+\bar{n}_0}(\boldsymbol{\xi}_{\text{post}}, \boldsymbol{\Omega}_{\text{post}}, \boldsymbol{\Delta}_{\text{post}}, \boldsymbol{\gamma}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}})$ with $\boldsymbol{\xi}_{\text{post}}$, $\boldsymbol{\Omega}_{\text{post}}$, $\boldsymbol{\Delta}_{\text{post}}$, $\boldsymbol{\gamma}_{\text{post}}$ and $\boldsymbol{\Gamma}_{\text{post}}$ defined as in Theorem 1.2. This result follows directly from an adaptation of Corollary 4 in Durante (2019); see also Theorem 1 in Fasano & Durante (2022). In particular, replacing $\mathbf{D}$ with $\bar{\mathbf{X}}_0$ and $\mathbf{I}_n$ with $\bar{\boldsymbol{\Sigma}}_0$ in Corollary 4 by Durante (2019), under a $\text{SUN}_{\bar{p},\bar{n}}(\boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\Delta}_1, \boldsymbol{\gamma}_1, \boldsymbol{\Gamma}_1)$ prior, yields the expressions for $\boldsymbol{\xi}_{\text{post}}$, $\boldsymbol{\Omega}_{\text{post}}$, $\boldsymbol{\Delta}_{\text{post}}$ and $\boldsymbol{\Gamma}_{\text{post}}$ in Theorem 1.2. Inclusion of the offset $\bar{\mathbf{y}}_0$ within the proof of Corollary 4 by Durante (2019) poses no difficulties since it directly enters the SUN truncation parameter, thereby providing the expression for $\boldsymbol{\gamma}_{\text{post}}$ in Theorem 1.2. □

As discussed in Section 1.3.1, the availability of a closed–form SUN posterior in Theorem 1.2 facilitates Bayesian inference for the whole class of models in Sections 1.2.1–1.2.4, by leveraging known properties of SUNs (e.g., Azzalini & Capitanio, 2013; Arellano-Valle & Azzalini, 2021). For instance, the moment generating function of the posterior is

$$M(\mathbf{t}) = e^{\boldsymbol{\xi}_{\text{post}}^{\intercal}\mathbf{t}+0.5\mathbf{t}^{\intercal}\boldsymbol{\Omega}_{\text{post}}\mathbf{t}}\frac{\Phi_{\bar{n}+\bar{n}_0}(\boldsymbol{\gamma}_{\text{post}}+\boldsymbol{\Delta}_{\text{post}}^{\intercal}\boldsymbol{\omega}_{\text{post}}\mathbf{t};\boldsymbol{\Gamma}_{\text{post}})}{\Phi_{\bar{n}+\bar{n}_0}(\boldsymbol{\gamma}_{\text{post}};\boldsymbol{\Gamma}_{\text{post}})}, \quad \mathbf{t}\in\Re^{\bar{p}}, \tag{1.12}$$

and, therefore, closed–form expressions for relevant moments can be obtained from (1.12). In particular, applying the derivations of Azzalini & Bacchieri (2010) and Arellano-Valle & Azzalini (2021) to the SUN posterior in Theorem 1.2, yields the following expressions for $\mathbb{E}[\boldsymbol{\beta}\mid\mathbf{y}]$ and $\text{var}[\boldsymbol{\beta}\mid\mathbf{y}]$

$$\begin{aligned}\mathbb{E}[\boldsymbol{\beta}\mid\mathbf{y}] &= \boldsymbol{\xi}_{\text{post}}+\boldsymbol{\omega}_{\text{post}}\boldsymbol{\Delta}_{\text{post}}\boldsymbol{\psi},\\ \text{var}[\boldsymbol{\beta}\mid\mathbf{y}] &= \boldsymbol{\Omega}_{\text{post}}+\boldsymbol{\omega}_{\text{post}}\boldsymbol{\Delta}_{\text{post}}(\boldsymbol{\Psi}-\boldsymbol{\psi}\boldsymbol{\psi}^{\intercal})\boldsymbol{\Delta}_{\text{post}}^{\intercal}\boldsymbol{\omega}_{\text{post}},\end{aligned} \tag{1.13}$$

where $\boldsymbol{\psi}$ is a $(\bar{n}+\bar{n}_0)\times 1$ vector having entries

$$\psi_i = \phi(\gamma_{\text{post},i})\Phi_{\bar{n}+\bar{n}_0-1}(\boldsymbol{\gamma}_{\text{post},-i}-\boldsymbol{\Gamma}_{\text{post},-i}\gamma_{\text{post},i};\boldsymbol{\Gamma}_{\text{post},-i,-i}-\boldsymbol{\Gamma}_{\text{post},-i}\boldsymbol{\Gamma}_{\text{post},-i}^{\intercal})/\Phi_{\bar{n}+\bar{n}_0}(\boldsymbol{\gamma}_{\text{post}};\boldsymbol{\Gamma}_{\text{post}})$$

for $i=1,\ldots,\bar{n}+\bar{n}_0$, with $\gamma_{\text{post},i}$ and $\boldsymbol{\gamma}_{\text{post},-i}$ denoting the ith element of $\boldsymbol{\gamma}_{\text{post}}$ and the $(\bar{n}+\bar{n}_0-1)\times 1$ vector obtained by removing entry i in $\boldsymbol{\gamma}_{\text{post}}$, respectively, whereas $\boldsymbol{\Gamma}_{\text{post},-i}$ and $\boldsymbol{\Gamma}_{\text{post},-i,-i}$ are the ith column of $\boldsymbol{\Gamma}_{\text{post}}$ without entry i and the sub–matrix obtained by removing the ith row and column from $\boldsymbol{\Gamma}_{\text{post}}$, respectively. Analogously, $\boldsymbol{\Psi}$ is a $(\bar{n}+\bar{n}_0)\times(\bar{n}+\bar{n}_0)$ symmetric matrix involving the second–order derivatives of the cumulative distribution function term in (1.12); refer to Arellano-Valle & Azzalini (2021) for the exact expression of $\boldsymbol{\Psi}$ and of higher–order moments of the SUN distribution. These quantities can be also computed via Monte Carlo since

$$(\boldsymbol{\beta}\mid\mathbf{y}) \overset{\text{d}}{=} \boldsymbol{\xi}_{\text{post}}+\boldsymbol{\omega}_{\text{post}}(\mathbf{U}_0+\boldsymbol{\Delta}_{\text{post}}\boldsymbol{\Gamma}_{\text{post}}^{-1}\mathbf{U}_1), \quad \text{(where } \overset{\text{d}}{=} \text{ means equality in distribution)} \tag{1.14}$$

with $\mathbf{U}_0\sim \text{N}_{\bar{p}}(\mathbf{0},\bar{\boldsymbol{\Omega}}_{\text{post}}-\boldsymbol{\Delta}_{\text{post}}\boldsymbol{\Gamma}_{\text{post}}^{-1}\boldsymbol{\Delta}_{\text{post}}^{\intercal})$ and $\mathbf{U}_1\sim \text{TN}_{\bar{n}+\bar{n}_0}(-\boldsymbol{\gamma}_{\text{post}};\mathbf{0},\boldsymbol{\Gamma}_{\text{post}})$, where the random variable $\text{TN}_{\bar{n}+\bar{n}_0}(-\boldsymbol{\gamma}_{\text{post}};\mathbf{0},\boldsymbol{\Gamma}_{\text{post}})$ denotes an $(\bar{n}+\bar{n}_0)$–variate Gaussian with mean $\mathbf{0}$, covariance matrix $\boldsymbol{\Gamma}_{\text{post}}$ and truncation below $-\boldsymbol{\gamma}_{\text{post}}$. This additive construction has been first derived in Arellano-Valle & Azzalini (2006) and allows to generate independent and identically distributed values from the exact posterior via linear combinations of samples from $\bar{p}$–variate Gaussians and $(\bar{n}+\bar{n}_0)$–variate truncated normals, thus overcoming convergence and mixing issues of MCMC methods; see Section 1.4 for details.

Uncertainty quantification and calculation of credible intervals is instead facilitated by the availability of a closed–form expression for the SUN cumulative distribution function. Adapting Azzalini & Bacchieri (2010) and Arellano-Valle & Azzalini (2021), this is

$$\text{pr}(\boldsymbol{\beta}\leq\mathbf{b}\mid\mathbf{y}) = \frac{\Phi_{\bar{p}+(\bar{n}+\bar{n}_0)}([(\mathbf{b}-\boldsymbol{\xi}_{\text{post}})^{\intercal}\boldsymbol{\omega}_{\text{post}}^{-1},\boldsymbol{\gamma}_{\text{post}}^{\intercal}]^{\intercal};\tilde{\boldsymbol{\Omega}}_{\text{post}})}{\Phi_{\bar{n}+\bar{n}_0}(\boldsymbol{\gamma}_{\text{post}};\boldsymbol{\Gamma}_{\text{post}})}, \quad \mathbf{b}\in\Re^{\bar{p}}, \tag{1.15}$$

where $\tilde{\boldsymbol{\Omega}}_{\text{post}}$ is a matrix with blocks $\tilde{\boldsymbol{\Omega}}_{\text{post}[11]}=\bar{\boldsymbol{\Omega}}_{\text{post}}$, $\tilde{\boldsymbol{\Omega}}_{\text{post}[21]}=\tilde{\boldsymbol{\Omega}}_{\text{post}[12]}^{\intercal}=-\boldsymbol{\Delta}_{\text{post}}^{\intercal}$ and $\tilde{\boldsymbol{\Omega}}_{\text{post}[22]}=\boldsymbol{\Gamma}_{\text{post}}$.

Extending the results of Durante (2019), Fasano et al. (2021) and Benavoli et al. (2021) to the general setting under consideration, it is also possible to obtain the marginal likelihood as follows

$$p(\mathbf{y}) = c \cdot \phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\xi}; \bar{\boldsymbol{\Sigma}}_1 + \bar{\mathbf{X}}_1\boldsymbol{\Omega}\bar{\mathbf{X}}_1^\intercal)\frac{\Phi_{\bar{n}+\bar{n}_0}(\boldsymbol{\gamma}_{\text{post}}; \boldsymbol{\Gamma}_{\text{post}})}{\Phi_{\bar{n}}(\boldsymbol{\gamma}; \boldsymbol{\Gamma})}, \tag{1.16}$$

where $c = 1$ under all the routinely–used models in Section 1.2 which rely on Gaussian utilities, namely (1.2), (1.3), (1.5), (1.6), (1.7) and (1.9), whereas for those formulations based on skewed utilities, e.g., (1.4), (1.8) and (1.10), the constant c is a known value. Albeit the primary interest is inference on $\boldsymbol{\beta}$, it is worth highlighting that the availability of the marginal likelihood in (1.16) allows to obtain empirical Bayes estimates for the other quantities in likelihood (1.1), such as the parameters in the covariance matrices $\bar{\boldsymbol{\Sigma}}_1$ and $\bar{\boldsymbol{\Sigma}}_0$, via numerical maximization; see Section 1.6 for further discussion on estimation of $\bar{\boldsymbol{\Sigma}}_1$ and $\bar{\boldsymbol{\Sigma}}_0$. In addition, equation (1.16) facilitates direct calculation of Bayes factors for model selection and evaluation of predictive probabilities. This second objective can be readily accomplished by noting that the predictive probability $p(\mathbf{y}_{\text{new}} \mid \mathbf{y})$ for a new vector of observations $\mathbf{y}_{\text{new}}$ from model (1.1) is equal to the ratio $p(\mathbf{y}_{\text{new}}, \mathbf{y})/p(\mathbf{y})$ of the two associated marginal likelihoods. Therefore, focusing for simplicity on the case $c = 1$ — which covers the most widely–used models in Section 1.2 — direct application of (1.16) yields

$$p(\mathbf{y}_{\text{new}} \mid \mathbf{y}) = \frac{\phi_{\bar{n}_1+\bar{n}_1^{\text{new}}}(\bar{\mathbf{y}}_{1\text{pred}} - \bar{\mathbf{X}}_{1\text{pred}}\boldsymbol{\xi}; \bar{\boldsymbol{\Sigma}}_{1\text{pred}} + \bar{\mathbf{X}}_{1\text{pred}}\boldsymbol{\Omega}\bar{\mathbf{X}}_{1\text{pred}}^\intercal)}{\phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\xi}; \bar{\boldsymbol{\Sigma}}_1 + \bar{\mathbf{X}}_1\boldsymbol{\Omega}\bar{\mathbf{X}}_1^\intercal)} \frac{\Phi_{\bar{n}+\bar{n}_0+\bar{n}_0^{\text{new}}}(\boldsymbol{\gamma}_{\text{pred}}; \boldsymbol{\Gamma}_{\text{pred}})}{\Phi_{\bar{n}+\bar{n}_0}(\boldsymbol{\gamma}_{\text{post}}; \boldsymbol{\Gamma}_{\text{post}})}. \tag{1.17}$$

where $\bar{n}_1^{\text{new}}$ and $\bar{n}_0^{\text{new}}$ are the dimensions of the vectors $\bar{\mathbf{y}}_{1\text{new}}$ and $\bar{\mathbf{y}}_{0\text{new}}$ associated with $\mathbf{y}_{\text{new}}$. Similarly, $\bar{\mathbf{y}}_{1\text{pred}} = (\bar{\mathbf{y}}_1^\intercal, \bar{\mathbf{y}}_{1\text{new}}^\intercal)^\intercal$, $\bar{\mathbf{X}}_{1\text{pred}} = (\bar{\mathbf{X}}_1^\intercal, \bar{\mathbf{X}}_{1\text{new}}^\intercal)^\intercal$ and $\bar{\boldsymbol{\Sigma}}_{1\text{pred}}$ denotes a block–diagonal matrix with $\bar{\boldsymbol{\Sigma}}_{1\text{pred}[11]} = \bar{\boldsymbol{\Sigma}}_1$ and $\bar{\boldsymbol{\Sigma}}_{1\text{pred}[22]} = \bar{\boldsymbol{\Sigma}}_{1\text{new}}$. The quantities $\boldsymbol{\gamma}_{\text{pred}}$ and $\boldsymbol{\Gamma}_{\text{pred}}$, and, implicitly, $\boldsymbol{\xi}_{\text{pred}}$, $\boldsymbol{\Omega}_{\text{pred}}$ and $\boldsymbol{\Delta}_{\text{pred}}$, are constructed analogously to the posterior parameters in Theorem 1.2, after replacing the original data with the enriched ones $(\bar{\mathbf{y}}_{1\text{pred}}, \bar{\mathbf{X}}_{1\text{pred}}, \bar{\boldsymbol{\Sigma}}_{1\text{pred}})$ and $(\bar{\mathbf{y}}_{0\text{pred}}, \bar{\mathbf{X}}_{0\text{pred}}, \bar{\boldsymbol{\Sigma}}_{0\text{pred}})$, where $\bar{\mathbf{y}}_{0\text{pred}} = (\bar{\mathbf{y}}_0^\intercal, \bar{\mathbf{y}}_{0\text{new}}^\intercal)^\intercal$, $\bar{\mathbf{X}}_{0\text{pred}} = (\bar{\mathbf{X}}_0^\intercal, \bar{\mathbf{X}}_{0\text{new}}^\intercal)^\intercal$ and $\bar{\boldsymbol{\Sigma}}_{0\text{pred}}$ is a block–diagonal matrix with $\bar{\boldsymbol{\Sigma}}_{0\text{pred}[11]} = \bar{\boldsymbol{\Sigma}}_0$ and $\bar{\boldsymbol{\Sigma}}_{0\text{pred}[22]} = \bar{\boldsymbol{\Sigma}}_{0\text{new}}$.

Before concluding the overview of the SUN properties that can facilitate posterior inference, it shall be emphasized that these variables are closed under marginalization, linear combinations and conditioning (Arellano-Valle & Azzalini, 2021). This means, for example, that the posterior distribution of each sub–vector $\boldsymbol{\beta}_{[\mathbf{j}]}$, $\mathbf{j} \subset \{1; \ldots; \bar{p}\}$ is $\text{SUN}_{|\mathbf{j}|,\bar{n}+\bar{n}_0}(\boldsymbol{\xi}_{\text{post}[\mathbf{j}]}, \boldsymbol{\Omega}_{\text{post}[\mathbf{jj}]}, \boldsymbol{\Delta}_{\text{post}[\mathbf{j}]}, \boldsymbol{\gamma}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}})$, where $\boldsymbol{\Delta}_{\text{post}[\mathbf{j}]}$ corresponds to the matrix $\boldsymbol{\Delta}_{\text{post}}$ after deleting all the rows whose indexes are not in $\mathbf{j}$. As a consequence, setting $\mathbf{j} = \{j\}$ shows that the posterior distribution of each regression coefficient β_j, $j = 1, \ldots, \bar{p}$ is still SUN. Similarly, the posterior distribution of the linear combination $\mathbf{a} + \mathbf{A}^\intercal\boldsymbol{\beta}$ is $\text{SUN}_{d,\bar{n}+\bar{n}_0}(\mathbf{a} + \mathbf{A}^\intercal\boldsymbol{\xi}_{\text{post}}, \mathbf{A}^\intercal\boldsymbol{\Omega}_{\text{post}}\mathbf{A}, [(\mathbf{A}^\intercal\boldsymbol{\Omega}_{\text{post}}\mathbf{A}) \odot \mathbf{I}_d]^{-1/2}\mathbf{A}^\intercal\boldsymbol{\omega}_{\text{post}}\boldsymbol{\Delta}_{\text{post}}, \boldsymbol{\gamma}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}})$. In particular, this implies that the posterior distribution of any linear predictor still belongs to the SUN family.

1.4 Computational methods

The results presented in Section 1.3.2 suggest that posterior inference under the broad class of models illustrated in Section 1.2 can be performed via closed–form solutions, without the need to rely on MCMC strategies or deterministic approximations. This is true for any, even huge, $\bar{p}$ as long as $\bar{n} + \bar{n}_0$ is small–to–moderate, but not when $\bar{n} + \bar{n}_0$ exceeds few hundreds (Durante, 2019; Fasano & Durante, 2022). In fact, equations (1.12)–(1.17) require evaluation of cumulative distribution functions of $(\bar{n} + \bar{n}_0)$–variate Gaussians or sampling from $(\bar{n} + \bar{n}_0)$–variate truncated normals which is known to be computationally challenging in high dimensions (Genz, 1992; Genz & Bretz, 2009; Chopin, 2011; Botev, 2017; Genton et al., 2018; Cao et al., 2019, 2021). This motivates still active research on developing sampling–based methods and accurate deterministic approximations for tractable Bayesian inference under the models presented in Section 1.2. Sections 1.4.1–1.4.3 review, unify, extend and compare both past and more recent developments along these lines.

1.4.1 Analytical methods

As discussed above, the evaluation of high–dimensional Gaussian integrals with linear constraints, such as those arising in equations (1.12)–(1.17), is a longstanding problem (e.g., Genz, 1992; Genz & Bretz, 2002; Miwa et al., 2003; Gassmann, 2003; Genz, 2004; Craig, 2008; Chopin, 2011; Hayter & Lin, 2013; Pakman & Paninski, 2014; Trinh & Genz, 2015; Nomura, 2016; Ridgway, 2016; Botev, 2017; Genton et al., 2018; Cao et al., 2019; Gessner et al., 2020; Cao et al., 2021).

A popular class of strategies for evaluating these Gaussian integrals encompasses several extensions of the original separation of variables estimator initially proposed by Genz (1992). This solution recasts the problem as a sequence of tractable one–dimensional integrals, which are evaluated numerically via a randomized quasi–Monte Carlo sampling. As suggested in, e.g., Genz & Bretz (2009), the variance of the resulting estimator can be further reduced by means of variable reordering. More recently, Botev (2017) proposed a new solution relying on an optimal exponential tilting of the Genz (1992) construction, which is found by solving efficiently a minimax saddle–point problem, and then used as proposal distribution of an importance sampler. While still providing an unbiased estimate, this technique achieves a rare vanishing asymptotic relative error property, which translates into a practical reduction of the estimator variance by orders of magnitude. Moreover, this enhanced procedure remains effective in settings where the original Genz (1992) method cannot provide reliable estimates. Such a solution, available in the `R` library `TruncatedNormal`, remains generally tractable in a few hundred of dimensions, but it progressively slows down beyond this regime. To achieve scalability in high dimensions, recent solutions leverage low–rank hierarchical block structures of the covariance matrix within the high–dimensional Gaussian integral to decompose the problem into a sequence of smaller–dimensional ones which facilitate reduction of computational complexity while

accounting for accuracy (Genton et al., 2018; Cao et al., 2019, 2021). Among these alternatives, the one proposed in Cao et al. (2021) provides a state–of–the–art extension of the original separation of variables estimator which incorporates both an effective tile–low–rank representation of the covariance matrix and an iterative block–reordering scheme to obtain notable improvements in runtimes and scalability. For instance, this solution has been recently adapted to the problem of evaluating predictive probabilities in high–dimensional probit Gaussian processes with $\bar{p}$ in tens of thousands (Cao et al., 2022), obtaining notable improvements over state–of–the–art methods.

There are also alternative solutions beyond the general separation of variables technique. For example, Pakman & Paninski (2014) proposed an Hamiltonian Monte Carlo scheme, incorporating the truncations via hard walls and exploiting the possibility to integrate exactly the Hamiltonian equations of motion with Gaussian potential. Ridgway (2016) developed a sequential Monte Carlo sampler for computing Gaussian orthant probabilities, adding a dimension at each step, combined with carefully–designed MCMC moves. More recently, Gessner et al. (2020) constructed an efficient estimator of Gaussian integrals with linear domain constraints, which decomposes the problem into a sequence of easier–to–solve conditional probabilities, relying on a sequence of nested domains. Each internal step uses an analytic version of elliptical slice sampling, exploiting the availability of closed–form solutions for the intersections between the ellipses and linear constraints. The authors reported evidence of the effectiveness of such method even for thousand–dimensional integrals. Additional relevant references for the problem of evaluating Gaussian integrals can be found in, e.g., Genz & Bretz (2002); Miwa et al. (2003); Gassmann (2003); Genz (2004); Craig (2008); Chopin (2011); Hayter & Lin (2013); Trinh & Genz (2015) and Nomura (2016).

Interestingly, some of the aforementioned strategies also provide, as a direct consequence, effective solutions for sampling from multivariate truncated normals, which can be useful to generate values from the SUN posterior via the additive representation in equation (1.14). Such methods can be found, for example, in Botev (2017) and Gessner et al. (2020). Motivated by inference on a phylogenetic multivariate probit model, Zhang et al. (2021b) and Nishimura et al. (2021) recently employed two alternative schemes for sampling from a truncated normal distribution with dimension above ten thousand. This is done by resorting to a bouncy particle sampler (BPS) and an Hamiltonian Zigzag sampler, respectively. The latter is a variant of HMC that rely on Laplace–distributed momentum. As such, it can be combined with the no–U–turn (NUTS) algorithm in Hoffman & Gelman (2014), benefiting from the associated minimal tuning.

All the above solutions provide effective methods for evaluating Gaussian cumulative distribution functions and, possibly, sampling from multivariate truncated normals. Nonetheless, such procedures are still subject to a tradeoff between accuracy and computational tractability which is often specific to the model analyzed and to the size of the data, thereby motivating still ongoing research. Due to this, it is difficult to find a generally–

applicable gold–standard among the aforementioned techniques, although, in practice, the method by Botev (2017) has often notable performance when applied to equations (1.12)–(1.17) in small–to–moderate settings with $\bar{n} + \bar{n}_0$ in the order of few hundreds. Higher–dimensional problems may require more scalable solutions (e.g., Cao et al., 2021; Gessner et al., 2020), even if more extensive empirical analyses are required to assess these methods in general settings.

1.4.2 Sampling–based methods

Whenever the interest is in more complex functionals of the SUN posterior distribution, beyond those presented in Section 1.3.2, an effective solution is to rely on Monte Carlo estimates based on samples from $p(\boldsymbol{\beta} \mid \mathbf{y})$. While generally–applicable MCMC strategies such as state–of–the–art implementations of Hamiltonian Monte Carlo (e.g., Hoffman & Gelman, 2014) and Metropolis–Hastings (e.g., Roberts & Rosenthal, 2001; Haario et al., 2001) can be considered, a widely–implemented class of algorithms in the context of the models presented in Section 1.2 are data augmentation Gibbs samplers (e.g., Chib, 1992; Albert & Chib, 1993; McCulloch & Rossi, 1994; Chib & Greenberg, 1998; Albert & Chib, 2001; Imai & Van Dyk, 2005; Holmes & Held, 2006). This is because most of the formulations discussed in Section 1.2 rely on Gaussian latent utilities which are assigned a regression model with coefficients $\boldsymbol{\beta}$. Therefore, treating these utilities as augmented data restores Gaussian–Gaussian conjugacy between the prior for $\boldsymbol{\beta}$ and the likelihood of the augmented utilities, which can be in turn sampled from truncated normal full–conditionals, given $\boldsymbol{\beta}$ and the censoring information provided by the observed $\mathbf{y}$. This facilitates the implementation of tractable Gibbs samplers that iterate among these two steps, thus producing samples from the posterior distribution of $\boldsymbol{\beta}$.

Although the above techniques have been proposed only for a subset of the models in Section 1.2, and in separate contributions mainly focusing on Gaussian priors (e.g., Chib, 1992; Albert & Chib, 1993; McCulloch & Rossi, 1994; Chib & Greenberg, 1998; Albert & Chib, 2001; Imai & Van Dyk, 2005; Holmes & Held, 2006), the comprehensive framework in equation (1.1), and the general conjugacy results reported in Section 1.3 allow to unify these MCMC strategies within a broad construction which can be applied to any model in Section 1.2, even beyond those currently studied, and holds not only for Gaussian priors, but also for general SUN ones. Letting $\mathbf{X}_{\text{post}} = \boldsymbol{\Delta}_{\text{post}}^{\intercal}\bar{\boldsymbol{\Omega}}_{\text{post}}^{-1}\bar{\boldsymbol{\omega}}_{\text{post}}^{-1}$, $\boldsymbol{\eta}_{\text{post}} = \boldsymbol{\gamma}_{\text{post}} - \mathbf{X}_{\text{post}}\boldsymbol{\xi}_{\text{post}}$ and $\boldsymbol{\Sigma}_{\text{post}} = \boldsymbol{\Gamma}_{\text{post}} - \boldsymbol{\Delta}_{\text{post}}^{\intercal}\bar{\boldsymbol{\Omega}}_{\text{post}}^{-1}\boldsymbol{\Delta}_{\text{post}}$, this general Gibbs sampler can be obtained by noticing that, due to (1.11), the kernel of the SUN posterior in Theorem 1.2 can be written as

$$
\begin{aligned}
p(\boldsymbol{\beta} \mid \mathbf{y}) &\propto \phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}; \boldsymbol{\Omega}_{\text{post}})\Phi_{\bar{n}+\bar{n}_0}(\boldsymbol{\gamma}_{\text{post}} + \boldsymbol{\Delta}_{\text{post}}^{\intercal}\bar{\boldsymbol{\Omega}}_{\text{post}}^{-1}\boldsymbol{\omega}_{\text{post}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}); \boldsymbol{\Gamma}_{\text{post}} - \boldsymbol{\Delta}_{\text{post}}^{\intercal}\bar{\boldsymbol{\Omega}}_{\text{post}}^{-1}\boldsymbol{\Delta}_{\text{post}}) \\
&\propto \phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}; \boldsymbol{\Omega}_{\text{post}}) \int \phi_{\bar{n}+\bar{n}_0}(\bar{\mathbf{z}} - (\boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\beta}); \boldsymbol{\Sigma}_{\text{post}})\mathbb{1}(\bar{\mathbf{z}} > \mathbf{0})\mathrm{d}\bar{\mathbf{z}} \\
&\propto \int p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y})\mathrm{d}\bar{\mathbf{z}}.
\end{aligned}
\tag{1.18}
$$

Therefore, extending similar derivations for multinomial probit by Fasano & Durante (2022) and leveraging standard properties of multivariate Gaussian and truncated normals, equation (1.18) implies a generally applicable data augmentation Gibbs sampler relying on the full–conditional distributions

$$\begin{aligned}(\boldsymbol{\beta} \mid \mathbf{y}, \bar{\mathbf{z}}) &\sim \mathrm{N}_{\bar{p}}(\mathbf{V}_{\text{post}}[\mathbf{X}_{\text{post}}^{\intercal}\boldsymbol{\Sigma}_{\text{post}}^{-1}(\bar{\mathbf{z}} - \boldsymbol{\eta}_{\text{post}}) + \boldsymbol{\Omega}_{\text{post}}^{-1}\boldsymbol{\xi}_{\text{post}}], \mathbf{V}_{\text{post}}), \\ (\bar{\mathbf{z}} \mid \mathbf{y}, \boldsymbol{\beta}) &\sim \mathrm{TN}_{\bar{n}+\bar{n}_0}(\mathbf{0}; \boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\text{post}}),\end{aligned} \tag{1.19}$$

where $\mathbf{V}_{\text{post}} = (\boldsymbol{\Omega}_{\text{post}}^{-1} + \mathbf{X}_{\text{post}}^{\intercal}\boldsymbol{\Sigma}_{\text{post}}^{-1}\mathbf{X}_{\text{post}}^{\intercal})^{-1}$. Hence, available Gibbs sampler for specific models within (1.1) and yet unexplored extensions to the whole class under general SUN priors, can be readily obtained as special cases of (1.19) under suitable specification of the posterior parameters defining the above full–conditionals. It shall also be emphasized that the sampling from the $(\bar{n} + \bar{n}_0)$–dimensional truncated normal distribution in (1.19) is usually simplified by the conditional independence properties among the latent utilities, underlying most of the models presented in Sections 1.2.1–1.2.4. This means that $\boldsymbol{\Sigma}_{\text{post}}$ is either diagonal or block–diagonal, often with small–dimensional blocks, and, therefore, sampling from $(\bar{\mathbf{z}} \mid \mathbf{y}, \boldsymbol{\beta})$ simply requires drawing values from univariate or low–dimensional truncated normals. Nonetheless, as discussed in Johndrow et al. (2018) the dependence structure between $\boldsymbol{\beta}$ and $\bar{\mathbf{z}}$ can still yield poorly–mixing implementations; see also Qin & Hobert (2019) for a detailed convergence analysis.

An effective option to obviate the above mixing issues is to sample i.i.d. values from the joint posterior $p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y})$, instead of autocorrelated ones as in scheme (1.19). Extending the derivations by Holmes & Held (2006) to the whole class of models in (1.1), under SUN priors (1.11), this task can be accomplished by noting that $p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}) = p(\boldsymbol{\beta} \mid \mathbf{y}, \bar{\mathbf{z}})p(\bar{\mathbf{z}} \mid \mathbf{y})$, where $p(\boldsymbol{\beta} \mid \mathbf{y}, \bar{\mathbf{z}})$ is the density of the Gaussian in (1.19), whereas $p(\bar{\mathbf{z}} \mid \mathbf{y})$ is obtained by marginalizing out from the truncated normal in (1.19) the $\boldsymbol{\beta}$ coefficients with density $\phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}; \boldsymbol{\Omega}_{\text{post}})$. Leveraging standard properties of Gaussian and truncated normal random variables, and recalling Holmes & Held (2006), this marginalization step implies

$$(\bar{\mathbf{z}} \mid \mathbf{y}) \sim \mathrm{TN}_{\bar{n}+\bar{n}_0}(\mathbf{0}; \boldsymbol{\gamma}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}}). \tag{1.20}$$

Replacing the full–conditional multivariate truncated normal in (1.19) with the one in (1.20), yields a scheme for sampling i.i.d. values from $p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y})$ and, as a direct consequence, from the posterior $p(\boldsymbol{\beta} \mid \mathbf{y})$ of interest. To do this, it is sufficient to draw $\bar{\mathbf{z}}$ from (1.20) and then generate a value for $\boldsymbol{\beta}$ by sampling from the Gaussian in (1.19) with mean evaluated at the sampled value of $\bar{\mathbf{z}}$. This routine is closely related to the i.i.d. sampler based on the additive representation of the SUN in equation (1.14) which relies on linear combination among samples from $\bar{p}$–variate Gaussian and $(\bar{n}+\bar{n}_0)$–variate truncated normals (Durante, 2019; Fasano & Durante, 2022; Fasano et al., 2021).

Although the above strategies effectively address the potential mixing and convergence issues of the Gibbs sampler in (1.19), the multivariate truncated normal in (1.20) is often

more challenging from a computational perspective relative to the one in (1.19). In fact, marginalizing out $\boldsymbol{\beta}$ in $\text{TN}_{\bar{n}+\bar{n}_0}(\mathbf{0}; \boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\text{post}})$ induces dependence among the latent utilities in $\bar{\mathbf{z}}$. This means that, unlike $\boldsymbol{\Sigma}_{\text{post}}$, the covariance matrix $\boldsymbol{\Gamma}_{\text{post}}$ of the truncated normal in (1.20) has no more a diagonal or block–diagonal structure and, hence, $p(\bar{\mathbf{z}} \mid \mathbf{y})$ does not factorize as the product of univariate or low–dimensional truncated normals as for $p(\bar{\mathbf{z}} \mid \mathbf{y}, \boldsymbol{\beta})$ in (1.19), making sampling from (1.20) much more challenging when $(\bar{n} + \bar{n}_0)$ is large. In the context of probit models, Holmes & Held (2006) address this issue by leveraging closure under conditioning properties of truncated normals (Horrace, 2005) to sample iteratively from the univariate truncated normal full–conditionals $p(\bar{z}_i \mid \bar{\mathbf{z}}_{-i}, \mathbf{y})$, for $i = 1, \dots, \bar{n} + \bar{n}_0$. However, this strategy implies a Gibbs–sampling routine which may be still subject to mixing issues. Alternatively, it is possible to sample directly from $p(\bar{\mathbf{z}} \mid \mathbf{y})$ in (1.20) leveraging the state–of–the–art schemes presented in Section 1.4.1 (e.g., Botev, 2017; Gessner et al., 2020). Nonetheless, as mentioned in Section 1.4.1, there is still lack of a generally–applicable gold standard for any size of $\bar{p}$ and $\bar{n} + \bar{n}_0$, thus motivating alternative solutions for inference in high dimension, beyond sampling–based schemes. A unified view of these alternative strategies, with a main focus on past and recent developments in variational Bayes and expectation–propagation, is provided in Section 1.4.3.

1.4.3 Deterministic approximation–based methods

Even resorting to state–of–the–art techniques, sampling from the posterior distribution can become prohibitive for high–dimensional datasets and large sample sizes (Chopin & Ridgway, 2017). In such scenarios, an effective solution is to consider deterministic approximations of the exact posterior. Sections 1.4.3–1.4.3 provide a unified treatment of classical and more recent VB (Blei et al., 2017) and EP (Minka, 2001) approximations which are widely–implemented solutions in the context of the models considered in this Chapter; see Chopin & Ridgway (2017) for a review of alternative methods, such as Laplace approximation and INLA (Rue et al., 2009).

Variational Bayes

VB solves a constrained optimization problem which aims at finding the approximating density that is the closest in Kullback–Leiber (KL) divergence (Kullback & Leibler, 1951) to the exact posterior, among all the densities within a pre–specified tractable family facilitating Bayesian inference. Recalling Blei et al. (2017), in the context of models admitting conditionally conjugate constructions with global parameters $\boldsymbol{\beta}$ and local augmented data $\bar{\mathbf{z}}$ — such as for the formulations in Section 1.2 — the solution of the optimization problem often benefits from taking $p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y})$ as the target density to be approximated, which in turn would yield an approximation for $p(\boldsymbol{\beta} \mid \mathbf{y})$ after marginalizing out $\bar{\mathbf{z}}$ (Girolami & Rogers, 2006; Consonni & Marin, 2007; Fasano et al., 2022; Fasano & Durante, 2022). As for the choice of the approximating family $\mathcal{Q}$, classical solutions (e.g., Girolami & Rogers, 2006;

Consonni & Marin, 2007) rely on mean–field assumptions (e.g., Blei et al., 2017) which can be generally expressed as $\mathcal{Q}_{\text{MF}} = \{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) : q(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q(\boldsymbol{\beta}) \prod_{c=1}^{C} q(\bar{\mathbf{z}}_c)\}$, where $\bar{\mathbf{z}}_1, \dots, \bar{\mathbf{z}}_C$ are distinct sub–vectors of $\bar{\mathbf{z}}$, such that $\bar{\mathbf{z}} = (\bar{\mathbf{z}}_1^\intercal, \dots, \bar{\mathbf{z}}_C^\intercal)^\intercal$. Note that the choice of how to factorize $q(\bar{\mathbf{z}})$ in C independent blocks is often guided by the dependence structures in $\bar{\mathbf{z}}$. For instance, in models relying on conditionally independent latent utilities in $\bar{\mathbf{z}}$, such as those in Section 1.2, it is common to factorize $q(\bar{\mathbf{z}})$ consistent with these conditionally independent sub–vectors. In fact, as illustrated in the context of probit (e.g., Consonni & Marin, 2007) and multinomial probit (e.g., Girolami & Rogers, 2006), even without assuming a specific factorization for $q(\bar{\mathbf{z}})$, i.e., $C = 1$, the optimum $q^*_{\text{MF}}(\bar{\mathbf{z}})$ within the class $\mathcal{Q}_{\text{MF}}$ would still factorize as $\prod_{c=1}^{C} q^*_{\text{MF}}(\bar{\mathbf{z}}_c)$, where $\bar{\mathbf{z}}_1, \dots, \bar{\mathbf{z}}_C$ correspond to the subsets of conditionally independent latent utilities, as implied by the assumed model and prior.

Summarizing the above discussion, the MF–VB solution can be formalized as

$$\begin{aligned} q^*_{\text{MF}}(\boldsymbol{\beta}, \bar{\mathbf{z}}) &= \text{argmin}_{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) \in \mathcal{Q}_{\text{MF}}} \text{KL}[q(\boldsymbol{\beta}, \bar{\mathbf{z}}) || p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y})] \\ &= \text{argmax}_{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) \in \mathcal{Q}_{\text{MF}}} \text{ELBO}[q(\boldsymbol{\beta}, \bar{\mathbf{z}})] \\ &= \text{argmax}_{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) \in \mathcal{Q}_{\text{MF}}} \{-\text{KL}[q(\boldsymbol{\beta}, \bar{\mathbf{z}}) || p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y})] + \log p(\mathbf{y})\}, \end{aligned} \tag{1.21}$$

with $\mathcal{Q}_{\text{MF}} = \{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) : q(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q(\boldsymbol{\beta}) \prod_{c=1}^{C} q(\bar{\mathbf{z}}_c)\}$. Recalling, e.g., Blei et al. (2017), this optimization problem can be solved via a simple coordinate ascent variational inference (CAVI) algorithm, which iteratively updates the solution of the approximating densities for both $\boldsymbol{\beta}$ and $\bar{\mathbf{z}}$ through the equations $q^{(t)}_{\text{MF}}(\boldsymbol{\beta}) \propto \exp\{\mathbb{E}_{\bar{\mathbf{z}}}[\log p(\boldsymbol{\beta} \mid \mathbf{y}, \bar{\mathbf{z}})]\}$ and $q^{(t)}_{\text{MF}}(\bar{\mathbf{z}}_c) \propto \exp\{\mathbb{E}_{(\boldsymbol{\beta}, \bar{\mathbf{z}}_{-c})}[\log p(\bar{\mathbf{z}}_c \mid \mathbf{y}, \boldsymbol{\beta}, \bar{\mathbf{z}}_{-c})]\}$, for $c = 1, \dots, C$, where $\bar{\mathbf{z}}_{-c}$ coincides with $\bar{\mathbf{z}}$ without sub–vector $\bar{\mathbf{z}}_c$, whereas the expectation is taken with respect to the most recent update of the variational density over the other conditioning variables. Replacing the full–conditional distributions in these expressions with those in equation (1.19), and leveraging closure under conditioning properties of multivariate truncated normals (Horrace, 2005), yields a general MF–VB approximation which extends Girolami & Rogers (2006) and Consonni & Marin (2007) to the whole class of models and priors in Section 1.2–1.3, and can be easily obtained via CAVI updates

$$\begin{aligned} &q^{(t)}_{\text{MF}}(\boldsymbol{\beta}) \propto \phi_{\bar{p}}(\boldsymbol{\beta} - \mathbf{V}_{\text{post}}[\mathbf{X}^\intercal_{\text{post}} \boldsymbol{\Sigma}^{-1}_{\text{post}} (\mathbb{E}_{\bar{\mathbf{z}}}[\bar{\mathbf{z}}] - \boldsymbol{\eta}_{\text{post}}) + \bar{\boldsymbol{\Omega}}^{-1}_{\text{post}} \boldsymbol{\xi}_{\text{post}}]; \mathbf{V}_{\text{post}}), \\ &q^{(t)}_{\text{MF}}(\bar{\mathbf{z}}_c) \propto \phi_{n_c}(\bar{\mathbf{z}}_c - \mathbb{E}_{\boldsymbol{\beta}, \bar{\mathbf{z}}_{-c}}[\boldsymbol{\mu}_c]; \boldsymbol{\Sigma}_{\text{post}[c,c]} - \boldsymbol{\Sigma}_{\text{post}[c,-c]} (\boldsymbol{\Sigma}_{\text{post}[-c,-c]})^{-1} \boldsymbol{\Sigma}_{\text{post}[-c,c]}) \mathbb{1}(\bar{\mathbf{z}}_c > \mathbf{0}), \end{aligned} \tag{1.22}$$

for $c = 1, \dots, C$, where n_c corresponds to the dimension of the sub–vector $\bar{\mathbf{z}}_c$, whereas $\mathbb{E}_{\bar{\mathbf{z}}}[\bar{\mathbf{z}}] = (\mathbb{E}^\intercal_{\bar{\mathbf{z}}_1}[\bar{\mathbf{z}}_1], \dots, \mathbb{E}^\intercal_{\bar{\mathbf{z}}_C}[\bar{\mathbf{z}}_C])^\intercal$ and

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\beta}, \bar{\mathbf{z}}_{-c}}[\boldsymbol{\mu}_c] = \boldsymbol{\eta}_{\text{post}[c]} + \mathbf{X}_{\text{post}[c]} \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{\beta}] + \\ + \boldsymbol{\Sigma}_{\text{post}[c,-c]} (\boldsymbol{\Sigma}_{\text{post}[-c,-c]})^{-1} (\mathbb{E}_{\bar{\mathbf{z}}_{-c}}[\bar{\mathbf{z}}_{-c}] - \boldsymbol{\eta}_{\text{post}[-c]} - \mathbf{X}_{\text{post}[-c]} \mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{\beta}]). \end{aligned}$$

In equation (1.22) the quantities $\boldsymbol{\Sigma}_{\text{post}[c,c]}$, $\boldsymbol{\Sigma}_{\text{post}[c,-c]}$, $\boldsymbol{\Sigma}_{\text{post}[-c,c]}$, and $\boldsymbol{\Sigma}_{\text{post}[c,c]}$, correspond to the

four blocks of $\boldsymbol{\Sigma}_{\text{post}}$ when partitioned to highlight the sub–vector $\bar{\mathbf{z}}_c$ against all the others in $\bar{\mathbf{z}}_{-c}$. Similarly, $\mathbf{X}_{\text{post}[c]}$, $\mathbf{X}_{\text{post}[-c]}$, $\boldsymbol{\eta}_{\text{post}[c]}$ and $\boldsymbol{\eta}_{\text{post}[-c]}$ denote the rows of $\mathbf{X}_{\text{post}}$ and $\boldsymbol{\eta}_{\text{post}}$ corresponding to $\bar{\mathbf{z}}_c$ and $\bar{\mathbf{z}}_{-c}$, respectively. Hence, according to (1.22), MF–VB for the whole class of models and priors in Section 1.2–1.3 can be implemented via simple CAVI routines providing Gaussian and truncated normal approximating densities for $\boldsymbol{\beta}$ and $\bar{\mathbf{z}}_1, \ldots, \bar{\mathbf{z}}_C$, respectively, which only require updating of the corresponding means with respect to the most recent density estimate of the other conditioning variables, until convergence of the ELBO. Computing the Gaussian expectation $\mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{\beta}]$ poses no computational difficulties, whereas, recalling Sections 1.3.2 and 1.4.1, evaluating the mean $\mathbb{E}_{\bar{\mathbf{z}}_c}[\bar{\mathbf{z}}_c]$, $c = 1, \ldots, C$ of the truncated normals may be challenging if n_c is large. Nonetheless, n_c is often equal to 1 or to a small value when factorizing $q(\bar{\mathbf{z}})$ consistent with the diagonal block structures of $\boldsymbol{\Sigma}_{\text{post}}$ that are implied by most of the models in Sections 1.2.1–1.2.4. This implies that the MF–VB solutions for the local variables $\bar{\mathbf{z}}$ correspond to tractable low–dimensional truncated normals whose expectation can be computed via efficient routines, such as the one in the R library `MomTrunc` (Galarza Morales et al., 2021).

Although MF–VB provides a scalable and widely–applicable solution under the models considered in this Chapter, as shown by Fasano et al. (2022) in the context of probit regression with Gaussian priors, the resulting Gaussian approximation $q^*_{\text{MF}}(\boldsymbol{\beta})$ has often low accuracy, both theoretically and empirically, in high dimensions, especially when $\bar{p} > \bar{n} + \bar{n}_0$. These issues are evident not only in a general underestimation of posterior uncertainty, but also in the tendency to overshrink locations and to induce bias in the predictive probabilities, thus affecting the reliability of Bayesian inference under $q^*_{\text{MF}}(\boldsymbol{\beta})$. To address these fundamental issues and improve the accuracy of VB in high dimensions, Fasano et al. (2022) and Fasano & Durante (2022) propose a partially–factorized MF–VB solution (PFM–VB) that replaces the classical mean–field family $\mathcal{Q}_{\text{MF}} = \{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) : q(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q(\boldsymbol{\beta}) \prod_{c=1}^{C} q(\bar{\mathbf{z}}_c)\}$ with a more flexible partially–factorized one $\mathcal{Q}_{\text{PFM}} = \{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) : q(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) \prod_{c=1}^{C} q(\bar{\mathbf{z}}_c)\}$ which avoids assuming independence between $\boldsymbol{\beta}$ and $\bar{\mathbf{z}}$ as in mean–field, and only factorizes $q(\bar{\mathbf{z}})$ as $\prod_{c=1}^{C} q(\bar{\mathbf{z}}_c)$. The structure of this enlarged family is directly motivated by the form of the actual joint posterior $p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y})$. In fact, as highlighted in Section 1.4.2, $p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y})$ can be re–written as $p(\boldsymbol{\beta} \mid \mathbf{y}, \bar{\mathbf{z}}) p(\bar{\mathbf{z}} \mid \mathbf{y})$, where $p(\boldsymbol{\beta} \mid \mathbf{y}, \bar{\mathbf{z}})$ is the density of the Gaussian full–conditional in (1.19), whereas $p(\bar{\mathbf{z}} \mid \mathbf{y})$ is the one of the $(\bar{n} + \bar{n}_0)$–variate truncated normal with full covariance matrix in (1.20); see also Holmes & Held (2006). Therefore, since the Gaussian form of $p(\boldsymbol{\beta} \mid \mathbf{y}, \bar{\mathbf{z}})$ does not seem to pose computational difficulties, it is reasonable to preserve dependence between $\boldsymbol{\beta}$ and $\bar{\mathbf{z}}$ in $\mathcal{Q}_{\text{PFM}}$ and only approximate the intractable multivariate truncated normal density $p(\bar{\mathbf{z}} \mid \mathbf{y})$ via the product $\prod_{c=1}^{C} q(\bar{\mathbf{z}}_c)$ of low–dimensional tractable ones. In addition, when the block partitions under MF–VB and PFM–VB coincide, $\mathcal{Q}_{\text{MF}} \subset \mathcal{Q}_{\text{PFM}}$. Hence, it is guaranteed that the optimum $q^*_{\text{PFM}}(\boldsymbol{\beta}, \bar{\mathbf{z}})$ under $\mathcal{Q}_{\text{PFM}}$ is never less accurate than $q^*_{\text{MF}}(\boldsymbol{\beta}, \bar{\mathbf{z}})$, namely $\text{KL}[q^*_{\text{PFM}}(\boldsymbol{\beta}, \bar{\mathbf{z}}) || p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y})] \leq \text{KL}[q^*_{\text{MF}}(\boldsymbol{\beta}, \bar{\mathbf{z}}) || p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y})]$.

The improved accuracy of the above procedure, combined with the simple solution

of the optimization problem even under the enlarged family $\mathcal{Q}_{\text{PFM}}$, have motivated subsequent extensions of the original idea in Fasano et al. (2022) to multinomial probit (Fasano & Durante, 2022), dynamic probit (Fasano & Rebaudo, 2021) and GPs (Cao et al., 2022), which can be, in fact, generalized to the whole class of models and priors in Sections 1.2–1.3. To clarify this result, note that by the chain rule of the KL divergence

$$\text{KL}[q(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) \prod_{c=1}^{C} q(\bar{\mathbf{z}}_c) || p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y})] = \mathbb{E}_{\bar{\mathbf{z}}}\big[\text{KL}[q(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) || p(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \mathbf{y})]\big] + \text{KL}[\prod_{c=1}^{C} q(\bar{\mathbf{z}}_c) || p(\bar{\mathbf{z}} \mid \mathbf{y})],$$

where the first non–negative summand is equal to zero only when $q(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) = p(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \mathbf{y})$. Hence, $q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}})$ coincides with the density of the exact Gaussian full–conditional distribution in (1.19), while the minimizer of $\text{KL}[\prod_{c=1}^{C} q(\bar{\mathbf{z}}_c) || p(\bar{\mathbf{z}} \mid \mathbf{y})]$ can be readily obtained by applying the closure under conditioning properties (Horrace, 2005) of the multivariate truncated normal in (1.20) to the CAVI equations $q_{\text{MF}}^{(t)}(\bar{\mathbf{z}}_c) \propto \exp\{\mathbb{E}_{\bar{\mathbf{z}}_{-c}}[\log p(\bar{\mathbf{z}}_c \mid \mathbf{y}, \bar{\mathbf{z}}_{-c})]\}$, $c = 1, \ldots, C$. These results yield the following scheme for obtaining $q^*_{\text{PFM}}(\boldsymbol{\beta}, \bar{\mathbf{z}})$, which replaces the one for the MF–VB solution in (1.22),

$$\begin{aligned} q_{\text{PFM}}(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) = p(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \mathbf{y}) &\propto \phi_{\bar{p}}(\boldsymbol{\beta} - \mathbf{V}_{\text{post}}[\mathbf{X}_{\text{post}}^{\intercal} \boldsymbol{\Sigma}_{\text{post}}^{-1}(\bar{\mathbf{z}} - \boldsymbol{\eta}_{\text{post}}) + \boldsymbol{\Omega}_{\text{post}}^{-1} \boldsymbol{\xi}_{\text{post}}]; \mathbf{V}_{\text{post}}), \\ q_{\text{PFM}}^{(t)}(\bar{\mathbf{z}}_c) &\propto \phi_{n_c}(\bar{\mathbf{z}}_c - \mathbb{E}_{\bar{\mathbf{z}}_{-c}}[\bar{\boldsymbol{\mu}}_c]; \boldsymbol{\Gamma}_{\text{post}[c,c]} - \boldsymbol{\Gamma}_{\text{post}[c,-c]}(\boldsymbol{\Gamma}_{\text{post}[-c,-c]})^{-1} \boldsymbol{\Gamma}_{\text{post}[-c,c]}) \mathbb{1}(\bar{\mathbf{z}}_c > \mathbf{0}), \end{aligned} \tag{1.23}$$

for $c = 1, \ldots, C$, with $\mathbb{E}_{\bar{\mathbf{z}}_{-c}}[\bar{\boldsymbol{\mu}}_c] = \boldsymbol{\gamma}_{\text{post}[c]} + \boldsymbol{\Gamma}_{\text{post}[c,-c]}(\boldsymbol{\Gamma}_{\text{post}[-c,-c]})^{-1}(\mathbb{E}_{\bar{\mathbf{z}}_{-c}}[\bar{\mathbf{z}}_{-c}] - \boldsymbol{\gamma}_{\text{post}[-c]})$, where the expectation is taken with respect to most recent density estimate of the conditioning variables, whereas the indexing of sub–vectors and matrix blocks is the same as the one detailed in equation (1.22).

As for the scheme of MF–VB in (1.22), also the CAVI for PFM–VB simply requires to update mean vectors until convergence of the ELBO. However, unlike for (1.22), this scheme is only required for the truncated normal components, whereas the solution for $q_{\text{PFM}}(\boldsymbol{\beta} \mid \bar{\mathbf{z}})$ is already known to coincide with $p(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \mathbf{y})$. This gain comes, however, at the cost that, unlike for MF–VB, the approximated density $q^*_{\text{PFM}}(\boldsymbol{\beta})$ of interest is not available as a direct output of (1.23). Recalling, Fasano et al. (2022) and Fasano & Durante (2022), this apparent drawback can be easily addressed after noticing that, by (1.23), $q^*_{\text{PFM}}(\boldsymbol{\beta} \mid \bar{\mathbf{z}})$ is the density of the random variable distributed as a linear combination between a Gaussian, with mean vector $\mathbf{V}_{\text{post}}(-\mathbf{X}_{\text{post}}^{\intercal} \boldsymbol{\Sigma}_{\text{post}}^{-1} \boldsymbol{\eta}_{\text{post}} + \bar{\boldsymbol{\Omega}}_{\text{post}}^{-1} \boldsymbol{\xi}_{\text{post}})$ and covariance matrix $\mathbf{V}_{\text{post}}$, and a random vector $\bar{\mathbf{z}}$ whose joint density is approximated via the product of low–dimensional truncated normals under the CAVI updates in (1.23). Recalling equation (1.14) this construction coincides with the additive representation of a $\text{SUN}_{\bar{p},\bar{n}+\bar{n}_0}$ random variable that, unlike for the exact SUN posterior in Theorem 1.2, relies on a block–diagonal matrix $\boldsymbol{\Gamma}_{\text{PFM}}$ with C low–dimensional $n_c \times n_c$ blocks, for $c = 1, \ldots, C$. This means that the computational challenges for closed–form inference under the exact SUN posterior discussed in Section 1.3.2 and 1.4.1 are no more present for the optimal SUN approximate density $q^*_{\text{PFM}}(\boldsymbol{\beta})$, since the $(\bar{n} + \bar{n}_0)$–variate Gaussian cumulative distribution functions and truncated nor-

mals in equations (1.12)–(1.17) now factorize as C low–dimensional components that can be effectively evaluated when $n_1, \ldots, n_C$ are small–to–moderate. For example,

$$\begin{aligned} \mathbb{E}_{\text{PFM}}[\boldsymbol{\beta}] &= \mathbf{V}_{\text{post}}[\mathbf{X}_{\text{post}}^{\intercal}\boldsymbol{\Sigma}_{\text{post}}^{-1}(\mathbb{E}_{\text{PFM}}[\bar{\mathbf{z}}] - \boldsymbol{\eta}_{\text{post}}) + \bar{\boldsymbol{\Omega}}_{\text{post}}^{-1}\boldsymbol{\xi}_{\text{post}}], \\ \text{var}_{\text{PFM}}[\boldsymbol{\beta}] &= \mathbf{V}_{\text{post}} + \mathbf{V}_{\text{post}}\mathbf{X}_{\text{post}}^{\intercal}\boldsymbol{\Sigma}_{\text{post}}^{-1}\text{var}_{\text{PFM}}[\bar{\mathbf{z}}]\boldsymbol{\Sigma}_{\text{post}}^{-1}\mathbf{X}_{\text{post}}\mathbf{V}_{\text{post}}, \end{aligned} \tag{1.24}$$

where $\mathbb{E}_{\text{PFM}}[\bar{\mathbf{z}}] = (\mathbb{E}_{\text{PFM}}[\bar{\mathbf{z}}_1^{\intercal}], \ldots, \mathbb{E}_{\text{PFM}}[\bar{\mathbf{z}}_C^{\intercal}])^{\intercal}$ comprises the expectation of each low–dimensional sub–vector $\bar{\mathbf{z}}_c$, $c = 1, \ldots, C$ with respect to its optimal truncated normal approximating density, whereas $\text{var}_{\text{PFM}}[\bar{\mathbf{z}}]$ is a block–diagonal covariance matrix with generic block $\text{var}_{\text{PFM}}[\bar{\mathbf{z}}]_{[c,c]}$ denoting the covariance matrix of $\bar{\mathbf{z}}_c$ according to its optimal truncated normal approximation. As previously mentioned, each of these quantities can be effectively evaluated in small–to–moderate dimensions via, e.g., the `R` `MomTrunc` (Galarza Morales et al., 2021). Recalling Fasano et al. (2022), the computational complexity of the PFM–VB is the same as the one for MF–VB, although the new partially–factorized solution yields improved accuracy both in theory and in practice. For instance, the authors prove that, unlike for MF–VB, the KL divergence between the PFM–VB approximation and the exact posterior density goes to 0 as $\bar{p} \to \infty$ and sample size fixed, thereby providing accurate inference in high–dimensional settings at a much lower computational cost, than exact solutions.

Expectation–propagation

Expectation–propagation (Minka, 2001) provides another well–established procedure for constructing a global approximation $q_{\text{EP}}^*(\boldsymbol{\beta})$ of the posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y})$ (Chopin & Ridgway, 2017; Riihimäki et al., 2014; Vehtari et al., 2020), which often yields improved accuracy in practice, relative to the MF–VB solution. Contrary to the mean–field VB methods presented in Section 1.4.3 — which only impose factorized structures for the approximating densities without necessarily assuming a functional form — EP postulates that the target posterior density itself can be written as a product of factors, also referred to as sites, and then iteratively approximates each one with an element of a given family of distributions, typically Gaussian for continuous variables or multinomial for discrete ones. Moreover, in the EP scheme each update is driven by the minimization of a suitable reverse KL, instead of the forward KL as in VB. This operation tends to improve MF–VB accuracy (e.g., Chopin & Ridgway, 2017) and becomes particularly convenient when the approximating density $q_{\text{EP}}(\boldsymbol{\beta})$ belongs to the exponential family, since it simply requires suitable moment matching strategies between $q_{\text{EP}}(\boldsymbol{\beta})$ and $p(\boldsymbol{\beta} \mid \mathbf{y})$ (e.g., Vehtari et al., 2020; Bishop, 2006, Chapter 10).

Current implementations of EP for probit (Chopin & Ridgway, 2017) and multinomial probit (Riihimäki et al., 2014) suggest that these strategies may yield practical gains for the whole class of models in Section 1.2, thus motivating the development of a broadly–applicable unified EP scheme, which is unavailable to date. This Section aims at covering such a gap, while providing novel closed–form expressions for moment matching of Gaussian sites leveraging the SUN conjugacy in Section 1.3, which also yields additional

supporting arguments on the accuracy of EP for the models in Section 1.2.

To address this goal, first note that, although likelihood (1.1) is very general, all the relevant examples discussed in Section 1.2 admit a factorized form for the intractable term

$$\Phi_{\bar{n}_0}(\bar{\mathbf{y}}_0 + \bar{\mathbf{X}}_0\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_0) = \prod_{c=1}^{C} \Phi_{\bar{n}_c}(\bar{\mathbf{y}}_{0[c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_{0[c,c]}), \tag{1.25}$$

where $\bar{\mathbf{y}}_0 = (\bar{\mathbf{y}}_{0[1]}^\intercal, \ldots, \bar{\mathbf{y}}_{0[C]}^\intercal)^\intercal$, $\bar{\mathbf{X}}_0 = (\bar{\mathbf{X}}_{0[1]}^\intercal, \ldots, \bar{\mathbf{X}}_{0[C]}^\intercal)^\intercal$ and $\bar{\boldsymbol{\Sigma}}_0$ is a block–diagonal matrix with generic block $\bar{\boldsymbol{\Sigma}}_{0[c,c]}$, for $c = 1, \ldots, C$. As discussed in Sections 1.4.2 and 1.4.3, this factorization is implied by the conditional independence structures among latent utilities, which yield tractable one–dimensional (e.g., probit and tobit) or low–dimensional (e.g., multinomial probit) factors $\Phi_{\bar{n}_c}(\bar{\mathbf{y}}_{0[c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\beta}, \bar{\boldsymbol{\Sigma}}_{0[c,c]})$. Hence, under these models, likelihood (1.1) can be written as

$$p(\mathbf{y} \mid \boldsymbol{\beta}) = p(\bar{\mathbf{y}}_1 \mid \boldsymbol{\beta})p(\bar{\mathbf{y}}_0 \mid \boldsymbol{\beta}) = \phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_1) \prod_{c=1}^{C} \Phi_{\bar{n}_c}(\bar{\mathbf{y}}_{0[c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_{0[c,c]}), \tag{1.26}$$

thus providing a general factorized structure which motivates EP schemes. For ease of notation, this routine is derived below under a Gaussian prior $p(\boldsymbol{\beta}) = \phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega})$ instead of its SUN generalization, although the proposed EP scheme can be easily extended to any prior within the SUN class. Updating this Gaussian prior with the likelihood in (1.26) yields the posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y})$ which can be more conveniently re–expressed as

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{y}) &= \frac{1}{p(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1)} \frac{p(\boldsymbol{\beta})p(\bar{\mathbf{y}}_1 \mid \boldsymbol{\beta})}{p(\bar{\mathbf{y}}_1)} p(\bar{\mathbf{y}}_0 \mid \boldsymbol{\beta}) = \frac{1}{p(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1)} p(\boldsymbol{\beta} \mid \bar{\mathbf{y}}_1)p(\bar{\mathbf{y}}_0 \mid \boldsymbol{\beta}), \\ &\propto \phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}; \boldsymbol{\Omega}_{\text{post}}) \textstyle\prod_{c=1}^{C} \Phi_{\bar{n}_c}(\bar{\mathbf{y}}_{0[c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_{0[c,c]}) \\ &= l_0(\boldsymbol{\beta}) \textstyle\prod_{c=1}^{C} l_c(\boldsymbol{\beta}) = \prod_{c=0}^{C} l_c(\boldsymbol{\beta}), \end{aligned} \tag{1.27}$$

where $l_c(\boldsymbol{\beta}) = \Phi_{\bar{n}_c}(\bar{\mathbf{y}}_{0[c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_{0[c,c]})$, $c = 1, \ldots, C$, correspond to the intractable terms in likelihood (1.26), whereas $l_0(\boldsymbol{\beta}) = p(\boldsymbol{\beta} \mid \bar{\mathbf{y}}_1) = \phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}; \boldsymbol{\Omega}_{\text{post}})$ is the conditional density obtained by updating the Gaussian prior $\phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega})$ for $\boldsymbol{\beta}$ with the tractable factor $\phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_1)$ in likelihood (1.26). As a direct consequence of the results in Section 1.3.2, this conditional density can be obtained in closed–form and coincides with the one of a Gaussian $\text{N}_{\bar{p}}(\boldsymbol{\xi}_{\text{post}}, \boldsymbol{\Omega}_{\text{post}})$, with parameters defined as in Theorem 1.2. Such a density acts as an intermediate prior in (1.27) to be updated with the intractable likelihood terms for obtaining the posterior $p(\boldsymbol{\beta} \mid \mathbf{y})$.

Recalling e.g., Vehtari et al. (2020), EP approximates the above posterior with a density $q_{\text{EP}}(\boldsymbol{\beta})$ that has the same factorized form of $p(\boldsymbol{\beta} \mid \mathbf{y})$ in (1.27), and is made of $C+1$ Gaussian sites. Hence

$$q_{\text{EP}}(\boldsymbol{\beta}) \propto \prod_{c=0}^{C} q_c(\boldsymbol{\beta}) = \prod_{c=0}^{C} \exp(-0.5\boldsymbol{\beta}^\intercal \mathbf{Q}_c \boldsymbol{\beta} + \boldsymbol{\beta}^\intercal \mathbf{r}_c) \\ = \exp[-0.5\boldsymbol{\beta}^\intercal \mathbf{Q}_{\text{EP}} \boldsymbol{\beta} + \boldsymbol{\beta}^\intercal \mathbf{r}_{\text{EP}}], \tag{1.28}$$

where $\mathbf{r}_c$ and $\mathbf{Q}_c$ define the natural parameters associated with the local Gaussian site c, for each $c = 1, \ldots, C$, whereas $\mathbf{r}_{\text{EP}} = \sum_{c=0}^{C} \mathbf{r}_c$ and $\mathbf{Q}_{\text{EP}} = \sum_{c=0}^{C} \mathbf{Q}_c$ denote those of the Gaussian EP approximation $q_{\text{EP}}(\boldsymbol{\beta})$ for $p(\boldsymbol{\beta} \mid \mathbf{y})$. Consistent with this expression, the ideal goal of EP would be to find the optimal $\mathbf{r}^*_{\text{EP}}$ and $\mathbf{Q}^*_{\text{EP}}$ such that the induced Gaussian density $q_{\text{EP}}(\boldsymbol{\beta})$ under (1.28) is as close as possible to the exact $p(\boldsymbol{\beta} \mid \mathbf{y})$ in (1.27) under the reverse KL divergence $\text{KL}[p(\boldsymbol{\beta} \mid \mathbf{y}) || q_{\text{EP}}(\boldsymbol{\beta})]$. Recalling Bishop (2006, Chapter 10), the solution of this optimization problem relies on a simple moment matching, which implies that $\mathbf{r}^*_{\text{EP}} = (\text{var}[\boldsymbol{\beta} \mid \mathbf{y}])^{-1}\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{y}]$ and $\mathbf{Q}^*_{\text{EP}} = (\text{var}[\boldsymbol{\beta} \mid \mathbf{y}])^{-1}$, or, alternatively, $\boldsymbol{\xi}^*_{\text{EP}} = \mathbb{E}[\boldsymbol{\beta} \mid \mathbf{y}]$ and $\boldsymbol{\Omega}^*_{\text{EP}} = \text{var}[\boldsymbol{\beta} \mid \mathbf{y}]$, where $\boldsymbol{\xi}^*_{\text{EP}}$ and $\boldsymbol{\Omega}^*_{\text{EP}}$ denote the mean vector and the covariance matrix of the Gaussian EP approximation. Unfortunately, as discussed in Section 1.3.2, the exact posterior is a SUN, and computing the associated moments is computationally challenging in general settings. In fact, such computational bottlenecks are those motivating the approximate schemes in Section 1.4.3.

To circumvent the aforementioned issue, EP relies on an iterative scheme which progressively improves $\mathbf{r}_{\text{EP}} = \sum_{c=0}^{C} \mathbf{r}_c$ and $\mathbf{Q}_{\text{EP}} = \sum_{c=0}^{C} \mathbf{Q}_c$ by sequentially updating each term $(\mathbf{r}_c, \mathbf{Q}_c)$, $c = 1, \ldots, C$, keeping fixed the others at their previous estimate (e.g., Vehtari et al., 2020). Let $l_{-c}(\boldsymbol{\beta}) = \prod_{c' \neq c} l_{c'}(\boldsymbol{\beta})$ and $q_{-c}(\boldsymbol{\beta}) = \prod_{c' \neq c} q_{c'}(\boldsymbol{\beta})$ denote the product among the different factors in (1.27) and (1.28), respectively, excluding the c–th one, this routine proceeds by optimizing, for every site c, a more tractable approximation of the reverse $\text{KL}[p(\boldsymbol{\beta} \mid \mathbf{y}) || q_{\text{EP}}(\boldsymbol{\beta})]$ in which the exact posterior $p(\boldsymbol{\beta} \mid \mathbf{y}) \propto l_{-c}(\boldsymbol{\beta}) l_c(\boldsymbol{\beta})$ is replaced by the hybrid one $p_h^{(t_c)}(\boldsymbol{\beta} \mid \mathbf{y}) \propto q_{-c}^{(t_c)}(\boldsymbol{\beta}) l_c(\boldsymbol{\beta})$, where t_c is the step of the algorithm which updates site c at iteration t. Using $q_{-c}^{(t_c)}(\boldsymbol{\beta})$ instead of $l_{-c}(\boldsymbol{\beta})$ yields a more tractable density since, by (1.28), the kernel of the so-called cavity distribution $q_{-c}^{(t_c)}(\boldsymbol{\beta})$ is that of a Gaussian with natural parameters $\mathbf{r}_{-c}^{(t_c)}$ and $\mathbf{Q}_{-c}^{(t_c)}$ corresponding to $\mathbf{r}_{\text{EP}} - \mathbf{r}_c$ and $\mathbf{Q}_{\text{EP}} - \mathbf{Q}_c$, respectively, when $\mathbf{r}_{\text{EP}}$, $\mathbf{r}_c$, $\mathbf{Q}_{\text{EP}}$ and $\mathbf{Q}_c$ are fixed at their most recent estimate. Hence, the only intractable term in the kernel of $p_h^{(t_c)}(\boldsymbol{\beta} \mid \mathbf{y})$ is $l_c(\boldsymbol{\beta}) = \Phi_{\bar{n}_c}(\bar{\mathbf{y}}_{0[c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\beta}, \bar{\boldsymbol{\Sigma}}_{0[c,c]})$. As a result, adapting derivations in Section 1.3, this hybrid density can be expressed as

$$p_h^{(t_c)}(\boldsymbol{\beta} \mid \mathbf{y}) \propto q_{-c}^{(t_c)}(\boldsymbol{\beta}) l_c(\boldsymbol{\beta}) \\ \propto \phi_{\bar{p}}\big(\boldsymbol{\beta} - (\mathbf{Q}_{-c}^{(t_c)})^{-1}\mathbf{r}_{-c}^{(t_c)}; (\mathbf{Q}_{-c}^{(t_c)})^{-1}\big)\Phi_{\bar{n}_c}\big(\bar{\mathbf{y}}_{0[c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\beta}; \bar{\boldsymbol{\Sigma}}_{0[c,c]}\big),$$

which implies that $p_h^{(t_c)}(\boldsymbol{\beta} \mid \mathbf{y})$ is the density of $\text{SUN}_{\bar{p},\bar{n}_c}(\boldsymbol{\xi}_c, \boldsymbol{\Omega}_c, \boldsymbol{\Delta}_c, \boldsymbol{\gamma}_c, \boldsymbol{\Gamma}_c)$ with parameters $\boldsymbol{\Omega}_c = (\mathbf{Q}_{-c}^{(t_c)})^{-1}$, $\boldsymbol{\xi}_c = (\mathbf{Q}_{-c}^{(t_c)})^{-1}\mathbf{r}_{-c}^{(t_c)}$, $\boldsymbol{\Delta}_c = \bar{\boldsymbol{\Omega}}_c \boldsymbol{\omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{s}_c^{-1}$, $\boldsymbol{\gamma}_c = \mathbf{s}_c^{-1}(\bar{\mathbf{y}}_{0[c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\xi}_c)$ and $\boldsymbol{\Gamma}_c = \mathbf{s}_c^{-1}(\bar{\boldsymbol{\Sigma}}_{0[c,c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal)\mathbf{s}_c^{-1}$, where $\mathbf{s}_c = ([\bar{\boldsymbol{\Sigma}}_{0[c,c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal] \odot \mathbf{I}_{\bar{n}_c})^{1/2}$. Therefore, unlike the exact SUN posterior, this hybrid SUN is much more tractable since the dimension of the cumulative distribution function term is $\bar{n}_c$, and not $\sum_{c=1}^{C} \bar{n}_c$ as in $p(\boldsymbol{\beta} \mid \mathbf{y})$, under Gaussian prior. In fact, as previously discussed, $\bar{n}_c$ is either equal to 1 or to a low value

under most of the models presented in Sections 1.2.1–1.2.4. This means that inference under the SUN with density $p_h^{(t_c)}(\boldsymbol{\beta} \mid \mathbf{y})$ can be performed via the closed–form expressions in Section 1.3.2, which can be effectively evaluated when $\bar{n}_c$ is small; see also Section 1.4.1. In particular, it is possible to compute the mean $\mathbb{E}^{(t_c)}[\boldsymbol{\beta} \mid \mathbf{y}]$ and variance $\text{var}^{(t_c)}[\boldsymbol{\beta} \mid \mathbf{y}]$ of $\boldsymbol{\beta}$ with respect to the hybrid density $p_h^{(t_c)}(\boldsymbol{\beta} \mid \mathbf{y})$ via expressions (1.13) evaluated at parameters $\boldsymbol{\xi}_c, \boldsymbol{\Omega}_c, \boldsymbol{\Delta}_c, \boldsymbol{\gamma}_c$ and $\boldsymbol{\Gamma}_c$. Alternatively, leveraging the additive representation of the SUN in (1.14), it follows that

$$\begin{aligned} \mathbb{E}^{(t_c)}[\boldsymbol{\beta} \mid \mathbf{y}] &= \boldsymbol{\xi}_c + \boldsymbol{\omega}_c \boldsymbol{\Delta}_c \boldsymbol{\Gamma}_c^{-1} \mathbb{E}[\mathbf{U}_{1c}], \\ \text{var}^{(t_c)}[\boldsymbol{\beta} \mid \mathbf{y}] &= \boldsymbol{\Omega}_c - \boldsymbol{\omega}_c \boldsymbol{\Delta}_c \boldsymbol{\Gamma}_c^{-1} \boldsymbol{\Delta}_c^\intercal \boldsymbol{\omega}_c + \boldsymbol{\omega}_c \boldsymbol{\Delta}_c \boldsymbol{\Gamma}_c^{-1} \text{var}[\mathbf{U}_{1c}] \boldsymbol{\Gamma}_c^{-1} \boldsymbol{\Delta}_c^\intercal \boldsymbol{\omega}_c, \end{aligned} \tag{1.29}$$

where $\mathbf{U}_{1c} \sim \text{TN}_{\bar{n}_c}(-\boldsymbol{\gamma}_c; \mathbf{0}, \boldsymbol{\Gamma}_c)$ is a low–dimensional truncated normal whose expectation $\mathbb{E}[\mathbf{U}_{1c}]$ and variance $\text{var}[\mathbf{U}_{1c}]$ can be effectively computed via R library `MomTrunc` (Galarza Morales et al., 2021), due to the small value of $\bar{n}_c$. This implies that the reverse KL can be easily optimized via moment matching when $p(\boldsymbol{\beta} \mid \mathbf{y})$ is replaced by $p_h^{(t_c)}(\boldsymbol{\beta} \mid \mathbf{y})$, thereby obtaining the improved estimates $\mathbf{r}_{\text{EP}}^{(t_c)}$ and $\mathbf{Q}_{\text{EP}}^{(t_c)}$ for the parameters of interest $\mathbf{r}_{\text{EP}}$ and $\mathbf{Q}_{\text{EP}}$ at step t_c, defined as $\mathbf{r}_{\text{EP}}^{(t_c)} = (\boldsymbol{\Omega}_{\text{EP}}^{(t_c)})^{-1} \boldsymbol{\xi}_{\text{EP}}^{(t_c)}$ and $\mathbf{Q}_{\text{EP}}^{(t_c)} = (\boldsymbol{\Omega}_{\text{EP}}^{(t_c)})^{-1}$, so that

$$\begin{aligned} \mathbf{r}_{\text{EP}}^{(t_c)} &= (\boldsymbol{\Omega}_c - \boldsymbol{\omega}_c \boldsymbol{\Delta}_c \boldsymbol{\Gamma}_c^{-1} \boldsymbol{\Delta}_c^\intercal \boldsymbol{\omega}_c + \boldsymbol{\omega}_c \boldsymbol{\Delta}_c \boldsymbol{\Gamma}_c^{-1} \text{var}[\mathbf{U}_{1c}] \boldsymbol{\Gamma}_c^{-1} \boldsymbol{\Delta}_c^\intercal \boldsymbol{\omega}_c)^{-1} (\boldsymbol{\xi}_c + \boldsymbol{\omega}_c \boldsymbol{\Delta}_c \boldsymbol{\Gamma}_c^{-1} \mathbb{E}[\mathbf{U}_{1c}]), \\ \mathbf{Q}_{\text{EP}}^{(t_c)} &= (\boldsymbol{\Omega}_c - \boldsymbol{\omega}_c \boldsymbol{\Delta}_c \boldsymbol{\Gamma}_c^{-1} \boldsymbol{\Delta}_c^\intercal \boldsymbol{\omega}_c + \boldsymbol{\omega}_c \boldsymbol{\Delta}_c \boldsymbol{\Gamma}_c^{-1} \text{var}[\mathbf{U}_{1c}] \boldsymbol{\Gamma}_c^{-1} \boldsymbol{\Delta}_c^\intercal \boldsymbol{\omega}_c)^{-1}. \end{aligned}$$

Concurrently, the new estimates of the parameters at site c — which are required for the subsequent updates — are $\mathbf{r}_c^{(t_c)} = \mathbf{r}_{\text{EP}}^{(t_c)} - \mathbf{r}_{-c}^{(t_c)}$ and $\mathbf{Q}_c^{(t_c)} = \mathbf{Q}_{\text{EP}}^{(t_c)} - \mathbf{Q}_{-c}^{(t_c)}$.

The above updating scheme is iterated multiple times $t \in \{1; \ldots\}$ and for each site $c = 1, \ldots, C$, until convergence to a stationary point. Note that in this routine site $c = 0$ does not require to be updated sequentially. Recalling, e.g. Chopin & Ridgway (2017) and Vehtari et al. (2020), factor $l_0(\boldsymbol{\beta})$ corresponds to the tractable Gaussian prior in (1.27) and, therefore, this term can be analytically matched to $q_0(\boldsymbol{\beta})$ in (1.28), obtaining $\mathbf{r}_0 = \boldsymbol{\Omega}_{\text{post}}^{-1} \boldsymbol{\xi}_{\text{post}}$ and $\mathbf{Q}_0 = \boldsymbol{\Omega}_{\text{post}}^{-1}$, where $\boldsymbol{\xi}_{\text{post}}$ and $\boldsymbol{\Omega}_{\text{post}}$ are defined as in Theorem 1.2. We shall also emphasize that the aforementioned EP scheme can yield, as a direct by–product, an approximation of the marginal likelihood $p(\mathbf{y})$. A detailed presentation of the step–by–step procedure to obtain such an estimate can be found in Appendix E of Vehtari et al. (2020), which shows that a key condition to compute such an approximation is the availability of the normalizing constant for the hybrid density $p_h^{(t_c)}(\boldsymbol{\beta} \mid \mathbf{y})$. Interestingly, this quantity is available in closed–form for the EP scheme discussed above since $p_h^{(t_c)}(\boldsymbol{\beta} \mid \mathbf{y})$ is the density of a $\text{SUN}_{\bar{p},\bar{n}_c}(\boldsymbol{\xi}_c, \boldsymbol{\Omega}_c, \boldsymbol{\Delta}_c, \boldsymbol{\gamma}_c, \boldsymbol{\Gamma}_c)$ and, hence, recalling Section 1.3, its normalizing constant is $\Phi_{\bar{n}_c}(\boldsymbol{\gamma}_c; \boldsymbol{\Gamma}_c)$. Since $\bar{n}_c$ is small, also this quantity can be effectively evaluated using, for example, the R library `TruncatedNormal` (Botev, 2017).

Although EP often yields improved accuracy relative to VB, it shall be noted that state–of–the–art implementations build on weaker theoretical guarantees relative to CAVI (e.g., Bishop, 2006; Chopin & Ridgway, 2017; Vehtari et al., 2020) and, as it will be discussed in

Section 1.4.3, tend to be more computationally demanding. For instance, there is no guarantee that the final EP solution minimizes the global reverse $\text{KL}[p(\boldsymbol{\beta} \mid \mathbf{y})||q_{\text{EP}}(\boldsymbol{\beta})]$, nor that the routine always converges in general settings. Nonetheless, empirical evidence typically reports remarkable EP accuracy in general settings, which is also confirmed by the simulation studies in Section 1.5. Recalling Bishop (2006, Chapther 10) an intuition for this notable performance is that, at each EP iteration the sites are updated to be most accurate in regions of high posterior probability controlled by the fixed remaining factors. More formal arguments can be found in Dehaene & Barthelmé (2015, 2018) and show that in asymptotic settings the discrepancy among the EP solution and the exact posterior goes to 0 faster than, for instance, the Laplace approximation. Recalling, e.g., Chopin & Ridgway (2017); Vehtari et al. (2020); Dehaene & Barthelmé (2015, 2018), these practical and theoretical accuracy results are intimately related to log–concavity properties of the target posterior. Interestingly, as shown recently in Arellano-Valle & Azzalini (2021, Section 3.1), SUN distributions are log–concave. Hence, the general conjugacy results in Section 1.3 are also useful in providing further support for the reliability of EP for the whole class of models presented in Section 1.2.

Computational costs

To conclude the analysis of the approximate schemes in Sections 1.4.3–1.4.3, we discuss the associated cost per–iteration focusing, for ease of notation, on classical probit regression $\prod_{i=1}^{n} \Phi(\mathbf{x}_i^\intercal \boldsymbol{\beta})^{\mathbb{1}(y_i=1)}[1-\Phi(\mathbf{x}_i^\intercal \boldsymbol{\beta})]^{\mathbb{1}(y_i=0)}$ as in (1.5), with a spherical Gaussian prior $p(\boldsymbol{\beta}) = \phi_{\bar{p}}(\boldsymbol{\beta}; \omega^2 \mathbf{I}_{\bar{p}})$. Note that, as discussed in Section 1.2.2, $n = \bar{n}_0$ and $p = \bar{p}$ when such a model is written as a special case of likelihood (1.1). Besides providing one of the most widely implemented formulations within the class of models whose likelihood can be expressed as in (1.1), this choice is also motivated by the fact that detailed costs per–iteration of effective MF–VB and PFM–VB implementations have been already derived in Fasano et al. (2022) under probit regression with Gaussian prior. Moreover, it gives the opportunity to show that currently–reported per–iteration costs of EP for the same class of models and priors (Chopin & Ridgway, 2017) can be further reduced, thus making also EP scalable to high dimensions.

For deriving the costs of MF–VB, PFM–VB and EP it shall be emphasized that, in probit regression, such approximations rely on $c = 1, \ldots, n$ and, hence, $n_c = \bar{n}_c = 1$ for any c. Under MF–VB (Consonni & Marin, 2007) and EP (Chopin & Ridgway, 2017), this choice is implied by the formulation of the optimization problem and is a direct consequence of the conditional independence among the unit–specific latent utilities. Instead, for PFM–VB (Fasano et al., 2022) such a setting is not enforced. Nonetheless, it provides a convenient specification which is in line with MF–VB and EP solutions, and also facilitates posterior inference by only requiring to deal with univariate truncated normals.

Under the above settings, Appendix A of Fasano et al. (2022) provides a detailed discussion of the per–iteration cost for both MF–VB and PFM–VB, which is $\mathcal{O}(n \cdot \min\{n; p\}) =$

$\mathcal{O}(\bar{n}_0 \cdot \min\{\bar{n}_0; \bar{p}\})$, after suitable matrix precomputations before running the CAVI routine. Since the mean and variance of univariate truncated normals can be accurately computed at $\mathcal{O}(1)$ cost under standard algorithms (e.g., Botev, 2017), the most intensive computations in the CAVI routines for MF–VB and PFM–VB are associated with the matrix multiplication operations. These steps can be efficiently implemented by exploiting recursive formulas when updating each univariate truncated normal approximating density conditioned the most recent estimate of the others, in PFM–VB, or of the $\boldsymbol{\beta}$ parameters, in MF–VB, thereby leading to an overall per–iteration cost that is either linear or sublinear in $p = \bar{p}$.

As for EP, the currently reported per–iteration cost in probit regression with spherical Gaussian priors is $\mathcal{O}(np^2) = \mathcal{O}(\bar{n}_0\bar{p}^2)$ (Chopin & Ridgway, 2017), after suitable precomputations as in MF–VB and PFM–VB. Intuitively, this increased complexity is due to the fact that, unlike for MF–VB and PFM–VB, not only the expectations but also the $p \times p$ covariance matrices must be updated and then inverted at every site c, for $c = 1, \ldots, n$. Although the specific form of such matrices allows to reduce the common cubic cost into a quadratic one via application of the Woodbury's formula to avoid direct matrix inversion (Chopin & Ridgway, 2017), an $\mathcal{O}(np^2) = \mathcal{O}(\bar{n}_0\bar{p}^2)$ cost is still computationally impractical in large $p = \bar{p}$ settings. In fact, as mentioned in the final discussion of Chopin & Ridgway (2017), even state–of–the–art implementations of EP are often computationally challenging when p exceeds one thousand. This is also confirmed in the empirical studies of Fasano et al. (2022), where the EP implementation within the R package `EPGLM` by Chopin & Ridgway (2017) requires more than six hours to reach convergence in a high–dimensional Alzheimer's application with $p = \bar{p} = 9036$ and $n = \bar{n}_0 = 300$. Notably, as a further contribution of the present Chapter, it shall be emphasized that a more scalable EP implementation with per–iteration cost $\mathcal{O}(np \cdot \min\{n; p\}) = \mathcal{O}(\bar{n}_0\bar{p} \cdot \min\{\bar{n}_0; \bar{p}\})$ can be actually derived by leveraging similar results considered in Fasano et al. (2022) for obtaining efficient implementations of MF–VB and PFM–VB. In particular, this novel EP implementation exploits the fact that, under the same reformulation via Woodbury's identity of Chopin & Ridgway (2017), the site updates do not necessarily require direct computation of the aforementioned $p \times p$ matrices, since such quantities enter via the inner product with the $p \times n$ design matrices $\bar{\mathbf{X}}_0^\intercal$. Hence, when $p = \bar{p}$ is large, it is more convenient to update this product directly, without storing, updating or multiplying any $p \times p$ matrix. This yields an $\mathcal{O}(np) = \mathcal{O}(\bar{n}_0\bar{p})$ cost for each site, and to an overall cost for the $n = \bar{n}_0$ site updates of $\mathcal{O}(n^2p) = \mathcal{O}(\bar{n}_0^2\bar{p})$. In high dimensional settings, when $p \gg n$, this linear cost in $p = \bar{p}$ yields massive computational gains relative to the original $\mathcal{O}(np^2) = \mathcal{O}(\bar{n}_0\bar{p}^2)$ cost of `EPGLM` in Chopin & Ridgway (2017). For example, applying the proposed more scalable implementation to the high–dimensional Alzheimer's application yields an overall runtime of five minutes, which is orders of magnitude lower than `EPGLM` (Chopin & Ridgway, 2017) that requires, instead, more than six hours. When, instead, $n \gg p$, the linear cost in $n = \bar{n}_0$ of the `EPGLM` ensures effective implementations. Combining these two scenarios yields an overall per–iteration cost of $\mathcal{O}(np \cdot \min\{n; p\}) = \mathcal{O}(\bar{n}_0\bar{p} \cdot \min\{\bar{n}_0; \bar{p}\})$, which is linear in the higher between $n = \bar{n}_0$ and

$p = \bar{p}_0$. To the best of our knowledge, this is the first implementation of EP available in the literature to achieve such computational efficiency. Despite this, standard EP remains more computationally demanding than MF–VB and PFM–VB since, as discussed above, also the $p \times p$ matrices need to be updated at each step of the EP routine, either directly or implicitly via the product with $\bar{\mathbf{X}}_0^\intercal$.

The above reasoning can be directly applied to highlight a similar dependence on sample size and number of predictors in the per–iteration cost of effective MF–VB, PFM–VB and EP implementations under the whole class of models and priors in Sections 1.2 and 1.3 — as long as n_c and $\bar{n}_c$ are sufficiently small to allow the calculations of the moments for the associated multivariate truncated normals at a negligible cost compared to the one of the matrix operations. This result is illustrated in empirical studies in Section 1.5, with a focus on tobit regression.

1.5 Empirical studies

Insightful empirical assessments of the methods in Sections 1.3–1.4, under specific models, can be found in Chopin & Ridgway (2017); Durante (2019); Fasano & Durante (2022); Cao et al. (2022); Fasano et al. (2022, 2021); Benavoli et al. (2020) and Benavoli et al. (2021); see also the Github repositories `ProbitSUN`, `Dynamic-Probit-PFMVB`, `Probit-PFMVB` and `PredProbitGP`. These studies encompass analyses of probit regression, multinomial probit, dynamic probit, probit Gaussian processes, skewed Gaussian processes and possible combinations of these constructions, but do not cover tobit regression for which SUN conjugacy has been proved in the present Chapter and, hence, the practical consequences of this results and the associated computational methods remain unexplored to date.

To address this key gap, we provide empirical evidence for the performance of the computational methods in Section 1.4, focusing on the standard tobit regression model in equation (1.9). In accomplishing this goal, we simulate a total of $n = n_0 + n_1 = 200$ observations from tobit regression, under three different proportions of censored observations $\mathtt{r} = n_0/n \in \{0.15; 0.50; 0.85\}$. This choice allows to cover a broad spectrum of scenarios which ranges from a model more similar to classical Gaussian linear regression, when $\mathtt{r} = 0.15$, to one closely mimicking unbalanced probit regression, when $\mathtt{r} = 0.85$. The p unit–specific predictors in $\mathbf{x}_i$, $i = 1, \ldots, n$, are instead simulated from standard Gaussians, except for the intercept term, whereas the regression coefficients in $\boldsymbol{\beta}$ are generated from a uniform distribution in the range $[-5, 5]$. Exploiting the latent utility interpretation of the tobit regression in Section 1.2, the final response data y_i, $i = 1, \ldots, n$ are obtained by first simulating the associated latent utilities z_i, $i = 1, \ldots, n$ from a $\mathrm{N}(\mathbf{x}_i^\intercal \boldsymbol{\beta}, 1)$, and then setting $y_i = z_i \mathbb{1}(z_i > z_{\mathrm{T}})$, for each $i = 1, \ldots, n$ where z_{T} is a pre–specified truncation threshold to obtain the desired proportion of censored observations under the three different settings of $\mathtt{r}$ considered. Note that this varying threshold poses no difficulties in Bayesian inference since it will directly enter the intercept term. To evaluate accuracy and computa-

Table 1.1: Runtimes, in seconds, of two alternative solutions to sample 5000 realizations from the posterior distribution in tobit regression with $n = 200$. (NUTS): `rstan` implementation of No–U–Turn HMC sampler. (i.i.d.): i.i.d. sampling from the exact SUN posterior via equation (1.14) leveraging the R library `TruncatedNormal`.

		p						
Censoring	*Method*	10	20	50	100	200	400	800
85 %	NUTS	3.18	3.97	11.71	24.88	56.48	150.62	1106.03
	i.i.d.	2.42	2.39	23.68	38.87	97.66	12.66	7.78
50 %	NUTS	2.47	2.59	3.90	10.32	42.32	119.60	1040.48
	i.i.d.	1.02	0.96	1.22	1.27	2.17	2.07	4.43
15 %	NUTS	1.65	1.79	2.60	4.77	18.97	90.91	460.38
	i.i.d.	0.11	0.09	0.12	0.18	0.39	0.91	3.02

tional efficiency at varying dimensions, these datasets are simulated for different values of $p \in \{10; 20; 50; 100; 200; 400; 800\}$. Posterior inference under the datasets produced for each combination of $\texttt{r}$ and p relies on spherical Gaussian priors $\text{N}_p(\mathbf{0}, \omega_p^2 \mathbf{I}_p)$, with $\omega_p^2 = 25 \cdot 10/p$ so as to control the variance of the linear predictor and induce increasing shrinkage in high dimensions. In addition, following recommended practice (e.g., Gelman et al., 2008; Chopin & Ridgway, 2017), the predictors are standardized to have 0 mean and standard deviation 0.5, before posterior inference.

Table 1.1 illustrates the computational gains in sampling–based methods which can be obtained by leveraging routines that exploit the SUN conjugacy in Section 1.3, instead of state–of–the–art MCMC alternatives. This is done by comparing, for every combination of $\texttt{r}$ and p, the runtimes to obtain 5000 samples from the exact posterior distribution of $\boldsymbol{\beta}$ under both the routinely–used **??**`rstan` implementation of No–U–Turn HMC sampler and the i.i.d. sampler which exploits the additive representation of the SUN posterior in (1.14). This latter routine leverages the R library `TruncatedNormal` (Botev, 2017) to sample the multivariate truncated normal component in (1.14). Consistent with related findings on probit (Durante, 2019) and multinomial probit (Fasano & Durante, 2022), Table 1.1 confirm the substantial computational gains of i.i.d. sampler relative to HMC in almost all settings of $\texttt{r}$ and p, especially when p is large. In fact, while high–dimensional regimes are often challenging for HMC, under (1.14) p only controls the dimension of the multivariate Gaussian which is feasible to sample, even in large p contexts. As discussed in Sections 1.4.1–1.4.2, more problematic for the i.i.d. scheme is the number of censored data n_0, which defines the dimension of the truncated normal in (1.14). This issue can be clearly seen in the increments of runtimes under i.i.d. sampling when the percentage of censoring grows from 15% to 85%. Nonetheless, the procedure remains still competitive relative to HMC in these small–to–moderate n_0 settings. It is also interesting to notice an increment in the runtime for the setting $\texttt{r} = 0.85$ (i.e., $n_0 = 170$), when $p \approx n_0$. In such a regime — which is reminiscent of the double–descent in high–dimensional regression (Hastie et al., 2022) — the

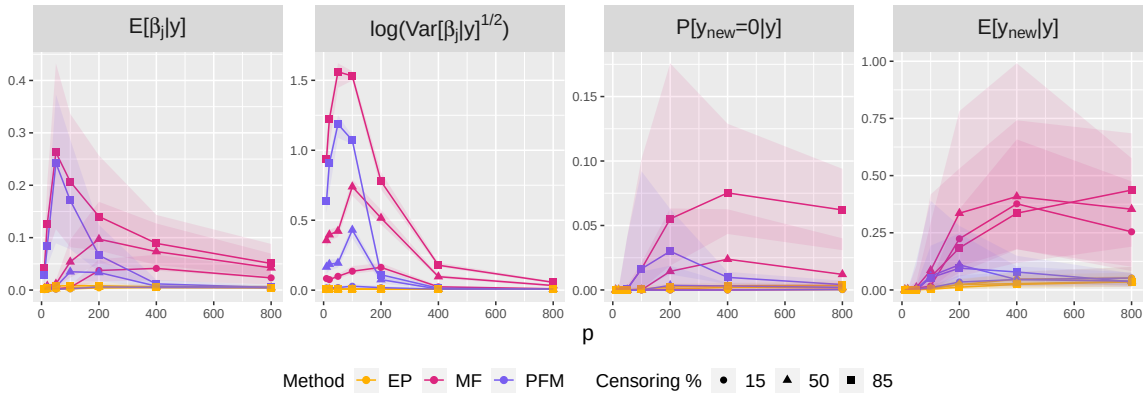

Figure 1.1: For four functionals of interest and different settings of $\mathtt{r} \in \{0.15; 0.50; 0.85\}$, trajectories for the median of the absolute differences, at varying p, between an accurate Monte Carlo estimate of such functionals via i.i.d. sampling from the exact SUN posterior and their approximation provided by mean–field variational Bayes (MF), partially–factorized variational Bayes (PFM) and expectation–propagation (EP) under tobit regression, with $n = 200$. The shaded areas correspond to the first and third quartiles computed from the absolute differences.

method by Botev (2017) experiences low acceptance probability that surely deserves further investigations.

As highlighted in Table 1.1, the moderate dimensions of the simulated datasets would still allow posterior inference under the closed–form solutions and i.i.d. sampling schemes presented in Section 1.3.2. Nonetheless, as already discussed, when n_0 grows, these procedures become computationally impractical, thereby motivating also the assessment of the more scalable approximate methods presented in Section 1.4.3. The relevant outcomes of these performance comparisons are reported in Figures 1.1–1.2 and in Table 1.2, with a focus on both accuracy and scalability. More specifically, Figure 1.1 provides insights on the accuracy of MF–VB, PFM–VB and EP in approximating key posterior functionals of interest at varying p, and for the three different settings of $\mathtt{r}$. These quantities include the posterior mean and variance of each β_j for $j = 1, \ldots, p$, along with predictive measures for the expected value of the response $\mathbb{E}[0 \cdot \Phi(-\mathbf{x}^\intercal_{\text{NEW},i}\boldsymbol{\beta}) + (\mathbf{x}^\intercal_{\text{NEW},i}\boldsymbol{\beta})\Phi(\mathbf{x}^\intercal_{\text{NEW},i}\boldsymbol{\beta}) \mid \mathbf{y}] = \mathbb{E}[(\mathbf{x}^\intercal_{\text{NEW},i}\boldsymbol{\beta})\Phi(\mathbf{x}^\intercal_{\text{NEW},i}\boldsymbol{\beta}) \mid \mathbf{y}]$ and the probability of a censoring event $\mathbb{E}[\Phi(-\mathbf{x}^\intercal_{\text{NEW},i}\boldsymbol{\beta}) \mid \mathbf{y}]$, both computed for 200 test observations whose predictors are simulated as for the original training data. For such functionals, Figure 1.1 displays medians and quartiles of the absolute differences between the corresponding Monte Carlo estimates under i.i.d. sampling from the exact posterior and the approximations provided by the three methods analyzed, at varying combinations of $\mathtt{r}$ and p. In the first two panels, the three quartiles are computed on the p absolute differences associated with coefficients $\beta_1, \ldots, \beta_p$, whereas in the last two panels these measures are calculated on the 200 absolute differences for the $i = 1, \ldots, 200$ test units.

Consistent with Chopin & Ridgway (2017) and despite the lack of theoretical guarantees, EP emerges as the most accurate solution in Figure 1.1 since its discrepancy from the

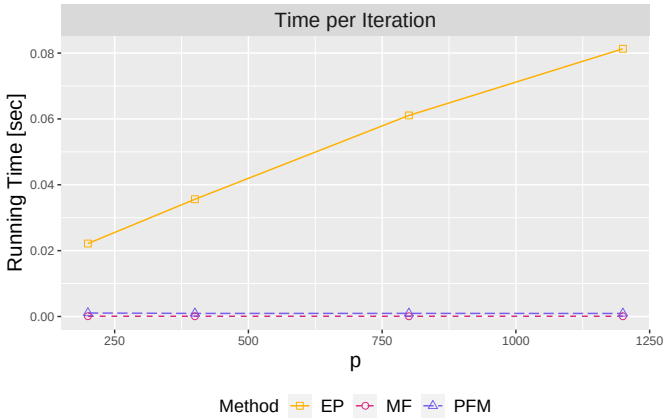

Figure 1.2: Runtime, in seconds, for each iteration of mean–field variational Bayes (MF), partially–factorized variational Bayes (PFM) and expectation–propagation (EP) in tobit regression at varying $p > n_0$ settings, when $\mathtt{r} = 0.5$.

Monte Carlo estimates is negligible in all regimes, ranging from $p \leq n_0$ to $p \geq n_0$. Nonetheless, as highlighted in Figure 1.2, the per–iteration runtimes of EP are linear in p for high dimensions, whereas those of MF–VB and PFM–VB are essentially constant and much lower. Therefore, from a computational perspective, such alternatives are more effective and scalable options in high–dimensional settings. This is especially true for PFM–VB which, as expected from the theory in Fasano et al. (2022), attains the same accuracy of EP when $p \gtrsim 2n_0$, but with a considerably lower computational effort. On the contrary, MF–VB is not competitive with EP in terms of accuracy and does not yield notable improvements in runtimes relative PFM–VB. Consistent with the theoretical results in Fasano et al. (2022), Table 1.2 provides evidence on the fact that the number of iterations needed by PFM–VB to reach convergence of the ELBO goes to one as p grows to infinity, while also displaying the

Table 1.2: Number of iterations required to reach convergence in mean–field variational Bayes (MF–VB), partially–factorized variational Bayes (PFM–VB) and expectation–propagation (EP) under tobit regression, with $n = 200$.

		p							
Censoring	*Method*	10	20	50	100	200	400	800	1200
85 %	MF–VB	192	407	337	342	180	147	92	103
	PFM–VB	126	271	178	127	14	7	4	4
	EP	4	5	7	6	5	4	4	4
50 %	MF–VB	39	67	75	138	192	146	149	172
	PFM–VB	22	37	34	42	17	6	3	4
	EP	4	3	4	4	5	4	4	4
15 %	MF–VB	11	16	20	36	90	90	114	169
	PFM–VB	7	9	9	11	10	3	4	4
	EP	3	3	3	3	4	4	3	3

phenomenon reminiscent of double–descent noticed in Table 1.1 for i.i.d. sampling, thus motivating further research along this line. Interestingly, the empirical results in Table 1.2 also suggest that the number of iterations required by EP does not grow with p. These analyses point toward EP as a default strategy, while suggesting PFM–VB as a valuable alternative in high–dimensional situations where EP is computationally impractical.

1.6 Discussion

This review Chapter provides a novel unified methodological and computational framework for Bayesian inference within a wide class of routinely–used regression models under a similarly broad set of prior distributions, which include the Gaussian one. Such an important gap in the literature is covered by first expressing the likelihoods associated with probit, tobit, multinomial probit and their extensions as special cases of a single formulation, and then generalizing early findings for specific models in e.g., Durante (2019); Fasano & Durante (2022); Fasano et al. (2021); Cao et al. (2022); Benavoli et al. (2020, 2021) to prove SUN conjugacy for any representation that admits such a general likelihood. This allows to develop general and broadly applicable versions of past and more recent computational methods, previously proposed only for some specific members of the general class and with a focus on Gaussian priors. These include data–augmentation Gibbs samplers, i.i.d. sampling schemes, mean–field variational Bayes, partially–factorized variational Bayes and novel scalable implementations of expectation–propagation.

Due to the relevance of the regression models considered within the present Chapter, such a review is expected to catalyze increasing interest by applied, computational and methodological researchers, and will hopefully motivate further research advancements along the directions opened by the results in Sections 1.2–1.5. For instance, the closed–form expressions in Section 1.3.2 for inference under the exact SUN posterior provide additional motivations to stimulate ongoing research aimed at developing accurate and fast methods to evaluate cumulative distribution functions of high–dimensional Gaussian distributions. In fact, any advancement along this direction and in sampling from multivariate truncated normals can be directly applied to conduct posterior inference via the closed–form results in Section 1.3.2, for increasingly larger sample sizes $\bar{n} + \bar{n}_0$, beyond small–to–moderate settings. This would be also useful for estimation of possible unknown parameters in the covariance matrices $\bar{\boldsymbol{\Sigma}}_1$ and $\bar{\boldsymbol{\Sigma}}_0$, via numerical maximization of the marginal likelihood $p(\mathbf{y})$ in (1.16). As shown in Sections 1.2.1–1.2.4 such matrices are often parameterized by a one–dimensional or low–dimensional vector of parameters, and hence can be effectively estimated via direct maximization of $p(\mathbf{y})$ when its evaluation is computationally practical. Alternatively, when the sample size exceeds small–to–moderate regimes, it is possible to optimize scalable approximations of $p(\mathbf{y})$, such as the one provided by EP. The availability of a closed–form expression (1.16) for $p(\mathbf{y})$ and of i.i.d. sampling schemes from $(\boldsymbol{\beta} \mid \mathbf{y})$ as in (1.14) can be also useful to improve full Bayesian inference for $\bar{\boldsymbol{\Sigma}}_1$ and $\bar{\boldsymbol{\Sigma}}_0$ when the

associated parameters are assigned a prior distribution. For example, leveraging $p(\mathbf{y})$ it is possible to derive collapsed Metropolis–Hastings routines to sample from the posteriors of $\bar{\boldsymbol{\Sigma}}_1$ and $\bar{\boldsymbol{\Sigma}}_0$ after integrating out $\boldsymbol{\beta}$ analytically, thereby improving mixing of MCMC schemes based on full–conditional distributions that rely both on $\boldsymbol{\beta}$ and on augmented data (e.g., Park & Van Dyk, 2009); see also Chan & Jeliazkov (2009) for effective MCMC methods to infer $\bar{\boldsymbol{\Sigma}}_0$ under identifiability constraints. These advancements are beyond the scope of this review, but provide a relevant research direction that is worth further exploration. Finally, it is also interesting to include hyperpriors for the scale parameters of the Gaussian or, more generally, SUN prior, that would yield scale–mixture representations which induce shrinkage in high dimensions (Carvalho et al., 2010). Since most of these constructions rely on conditionally Gaussian priors, the results in the present review may be useful to obtain improved theoretical and practical performance in state–of–the–art implementations of the models in Section 1.2 under sparse settings and more general classes of priors.

Although likelihood (1.1) already encompasses most of the models of interest in routine applications, further generalizations of such a likelihood and of the conjugacy results in Section 1.3 can be considered. For instance, it is possible to extend (1.1) to any version of the models in Sections 1.2.1–1.2.4 that arise from censoring or rounding of the Gaussian latent utilities into a generic truncation region. In fact, as discussed in Section 1.2.4, such a mechanism is directly related to the generative construction of the broader class of selection distributions (SLCT) in Arellano-Valle et al. (2006). Hence, following the same general reasoning considered in the present review, it seems natural to prove SLCT conjugacy for this broader family of likelihoods. For instance, this has been done in Kowal (2021) and King & Kowal (2021) by extending the ideas in Durante (2019) and Fasano et al. (2021) to static and dynamic rounded–data situations. These generalizations can be considered to prove similar conjugacy results for any extension of (1.1) which incorporates truncation into a finite region. Similarly, it would be also interesting to extend the recent conjugacy results under skew–elliptical link functions to the proposed general framework. In fact, likelihood (1.1) incorporates skew–normal and multivariate skew–normal utilities, but not generic skew–elliptical ones, such as skew–t. Motivated by results in Durante (2019) and Fasano & Durante (2022), Zhang et al. (2021a) prove that unified skew–elliptical distributions are conjugate to probit and multinomial probit with skew–elliptical link functions, thereby suggesting that such a result may hold more generally for any regression model whose induced likelihood arises from partially or fully observed skew–elliptical latent utilities.

Finally, it shall be emphasized that the class of models in Section 1.2 arguably encompasses the broadest set of formulations that appear in econometrics (Greene, 2008) and social sciences (DeMaris, 2004). Nonetheless, routine applications of such models under a Bayesian perspective have lagged behind the growing interest in Bayesian statistics. This is mainly due to the apparent intractability of posterior inference under such a class of regression models. This review not only clarifies that the posterior distributions induced by

the likelihoods of these models belong to a known class of variables, but also that such conjugacy results hold for a broader set of priors and for various extensions of classical probit, tobit and multinomial probit that are of direct relevance in econometrics and social science constructions. Therefore, the present Chapter will hopefully boost routine–use of these Bayesian models in applied research and motivate the development and implementation of even more flexible versions which still belong to likelihood (1.1), including, for example, random effect formulations and graphical models (e.g., Jones et al., 2005).

Chapter 2

Optimal lower bounds for logistic likelihoods

The use of logit mapping in binary regression models notoriously hinders tractable analytical inference. In the attempt to circumvent such difficulty, data-augmentation (DA) strategies for logistic regression have received considerable attention within the Bayesian framework. Conversely, unconstrained and penalized maximum likelihood (ML) estimations typically proceed via iterative schemes that alternate between the construction and the optimization of quadratic approximations of the logistic log-likelihood, either corresponding to Newton's method or arising from different tangent bounds exploited within minorize-maximize (MM) or expectation-maximization (EM) schemes. As Newton's method remains prone to unstable convergence issues, we focus our attention on the above strategies, giving new evidence on the optimality of the lower bound arising from the Pólya-Gamma data-augmentation scheme among quadratic minorizers for the logistic log-likelihood. We show that this advantage over alternative quadratic bounds is enhanced by the combination with ℓ_1-regularizations as a byproduct of the associated coordinate-wise optimization schemes. Furthermore, we derive a novel tangent minorizer dominating the Pólya-Gamma one, by adding a piece-wise linear contribution proportional to the ℓ_1-norm of the linear predictors. Such piece-wise quadratic bound still allows for a tractable coordinate-wise optimization algorithm, as routinely implemented in the literature for lasso and elastic net penalized logistic regression. Empirical results confirm that the higher flexibility of the proposed bound leads to an improved convergence rate of the resulting MM scheme.

2.1 Introduction

In the present Chapter, we turn our attention toward logistic regression, which models a set of binary observations $\mathbf{y} = (y_1, \dots, y_n)^\intercal$ as Bernoulli random variables, each with success probability $\pi(\mathbf{x}_i^\intercal \boldsymbol{\beta}) = (1 + e^{-\mathbf{x}_i^\intercal \boldsymbol{\beta}})^{-1}$, given a vector unknown parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\intercal \in \Re^p$ and a set of observed predictors $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\intercal \in \Re^p$, for $i =$

$1, \dots, n$. Despite such simple structure, the logit link notoriously hinders exact analytical inference, even from a Frequentist perspective. Accordingly, maximum likelihood (ML) estimation typically proceeds via the sequential optimization of quadratic approximations of the logistic log-likelihood, obviating the absence of a closed-form solution for the original problem (Hastie et al., 2015). The most natural quadratic approximation of the log-likelihood arises from a second-order truncation of Taylor expansion, as exploited within Newton-Raphson's method, chiefly appreciated because of its speed. However, it is well known that convexity of the objective function is not sufficient to guarantee convergence of the Newton-Raphson updates (McLachlan & Krishnan, 1996), which may incur into oscillating and even diverging behaviors (Böhning & Lindsay, 1988). Such unreliable convergence issues can be by solved resorting to minorize-maximize (Wu & Lange, 2010) and expectation-maximization schemes (McLachlan & Krishnan, 1996), that achieve a desirable monotonicity property by maximizing at every step tangent bounds for the target function. Indeed, the sequential optimization of such bounds is guaranteed to drive the log-likelihood uphill at each iteration, ensuring stability of the corresponding updates, potentially at the expense of slower convergence. For this reason, over the years several contributions focused on the construction and refinement of tractable tangent lower bounds for logistic log-likelihoods (Böhning & Lindsay, 1988; Jaakkola & Jordan, 2000; Marlin et al., 2011; Ermis & Bouchard, 2014).

A simple quadratic bound (BL) first appeared in the seminal work by Böhning & Lindsay (1988), by exploiting a uniform bound on the curvature of the logistic log-likelihood function to minorize the associated Hessian matrix, which has been largely implemented in the literature (Hunter & Lange, 2004; Wu & Lange, 2010; James, 2017; Khan et al., 2010). The same minorization strategy has been extended even to multinomial logistic regression (Böhning, 1992; Krishnapuram et al., 2005; Browne & McNicholas, 2015), as well as to penalized ML estimation (Friedman et al., 2007, 2010; Hastie et al., 2015), by combining the unaltered penalty term with the lower bound for the log-likelihood contribution. An alternative quadratic bound was derived in Jaakkola & Jordan (2000), by exploiting the supporting hyperplane inequality for a suitable transformation of the log-likelihood. Notwithstanding its good empirical performance and widespread use in the literature (Bishop & Svensén, 2003; Rasmussen & Williams, 2006; Lee et al., 2010; Ren et al., 2011; Carbonetto & Stephens, 2012), such routine apparently lacked a clear probabilistic interpretation. However, the recent contribution by Durante & Rigon (2019) provided an elegant justification for the procedure by Jaakkola & Jordan (2000), showing that the associated minorizer arises as proper evidence lower bound under the celebrated Pólya-Gamma (PG) data augmentation scheme for logistic regression (Polson et al., 2012). Accordingly, the optimization procedure by Jaakkola & Jordan (2000) can be regarded as a full-fledged EM scheme, in light of its connection with a well-defined missing variables representation. Durante & Rigon (2019) additionally carried out a direct comparison of the two MM schemes building on the BL and PG bound, showing that the routine leveraging on the latter dominates over

the former in terms of the asymptotic rates of convergence. Intuitively, this reflects how well the corresponding surrogates approximate the target function near a local extremum (Zhou & Zhang, 2012; Lange, 2016), suggesting that a tighter approximation will lead to a faster convergence rate. Indeed, following the reasoning by De Leeuw & Lange (2009), it can be shown that the PG bound is the sharpest tangent quadratic minorizer that can be constructed for the logistic log-likelihood. However, such optimality result is typically stated solely by restricting the attention to simpler one-dimensional building blocks of the likelihood function. Besides deriving an alternative proof for such univariate findings, in this Section 2.2 we reformulate and emphasize the optimality of the PG lower bound, by expressing the corresponding quadratic surrogate explicitly as a function of the regression parameter $\boldsymbol{\beta}$.

Nevertheless, the better converge rate associated with the PG bound comes with a trade-off in the efficiency of the corresponding updates for the joint optimization over $\Re^p$. In fact, while the fixed curvature of the BL bound allows to limit costly matrix inversion solely to the initialization of the procedure, the higher flexibility of the PG bound forces to repeat such calculations at each iteration. On the contrary, in Section 2.3 we highlight that the benefit coming from the sharpness of the bound becomes more evident in combination with some of the most popular regularization methods, such as the lasso (Tibshirani, 1996) and the elastic net (Zou & Hastie, 2005). Indeed, the ℓ_1-norm term in such penalties dictates the use of coordinate-wise descent schemes to solve the resulting optimization problem (Friedman et al., 2007, 2010), which indirectly avoids the aforementioned large matrix inversions. This implies that the costs per iteration associated with the BL and PG MM schemes coincide in the coordinate-wise optimization framework, limiting the relative performance to the difference between the corresponding convergence rates. Accordingly, the PG bound results to all extent preferable to the BL one in such penalized scenarios. Surprisingly the aforementioned computational advantage of the PG bound for penalized regression has been apparently overlooked even in state-the-art statistical software, albeit it might be readily incorporated with minor modifications.

Finally, in Section 2.4 we construct a novel piece-wise quadratic (PLQ) minorizer for logistic log-likelihoods, dominating over the PG bound. Indeed, several contributions in the literature focused on improving over the bound by Jaakkola & Jordan (2000) via piece-wise surrogates (Khan et al., 2010; Marlin et al., 2011; Ermis & Bouchard, 2014). For instance, Marlin et al. (2011) consider the general class of all piece-wise quadratic tangent bounds, defined on an arbitrary number of intervals. However, the flexibility of such formulations comes at the cost of reduced tractability, as the minorizers are defined implicitly and refined by data-agnostic numerical optimization. In contrast, the bound we propose allows for a simple analytical expression, as we complement the quadratic surrogate of the log-likelihood of each observation with a piece-wise linear contribution, proportional to the absolute value of the corresponding linear predictor $|\mathbf{x}_i^\intercal \boldsymbol{\beta}|$. Notably, the coordinate-wise updates for the PLQ bound still admit exact solutions, which make it particularly suitable for

the combination with the ℓ_1-penalties discussed above. The sharper approximation of the likelihood is expected to bring a further advantage in terms of the convergence rate of the associated MM optimization, compared to that of the BL bound. Indeed, in Section 2.5 we provide empirical evidence of the benefits brought by the proposed methodology, which becomes particularly appreciable in high-dimensional regression settings.

2.2 Minorize-maximize and expectation-maximization schemes

The acronym MM denotes a broad class of powerful iterative optimization schemes that address difficult minimization or maximization problems by solving a sequence of simpler surrogate ones, arising from the construction of tangent bounds for the original target function. The first appearances of the underlying principle trace back to De Leeuw (1977) and De Leeuw & Heiser (1977). However, except for the sub-class of EM algorithms, the more general formulation of MM schemes experienced a renewed popularity among the statistics community only in more recent years (Hunter & Lange, 2004; Wu & Lange, 2010; Lange et al., 2021), thanks to a beneficial combination of convergence guarantees and practical efficiency. In this Section, we concentrate on the minorize-maximize formulation of MM schemes, since the target optimization problem considered is that of maximum likelihood estimation for logistic regression, although the same MM acronym encompasses analogous minorize-maximize routines in case the interest is on maximization. Given a target function $\mathcal{F}(\boldsymbol{\beta})$ and a starting vector $\boldsymbol{\beta}^{(t)} \in \Re^p$, the technical essence of such methods lies in the construction of a tangent minorizer $\mathcal{G}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$, namely a function satisfying the properties

$$\begin{aligned} \mathcal{F}(\boldsymbol{\beta}^{(t)}) &= \mathcal{G}(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{\beta}^{(t)}) \\ \mathcal{F}(\boldsymbol{\beta}) &\geq \mathcal{G}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) \qquad \forall \boldsymbol{\beta} \in \Re^p. \end{aligned} \tag{2.1}$$

Although in principle there are no limitations on the analytic form of the lower bound, the above relation becomes of practical interest whenever the optimization of $\mathcal{G}$ is tractable and efficient, whereas that of $\mathcal{F}$ is troublesome or computationally demanding. Indeed, the optimization of the former via $\boldsymbol{\beta}^{(t+1)} = \text{argmax}_{\boldsymbol{\beta}} \mathcal{G}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ jointly drives uphill even the latter, as

$$\mathcal{F}(\boldsymbol{\beta}^{(t+1)}) \geq \mathcal{G}(\boldsymbol{\beta}^{(t+1)} \mid \boldsymbol{\beta}^{(t)}) \geq \mathcal{G}(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{\beta}^{(t)}) = \mathcal{F}(\boldsymbol{\beta}^{(t)}). \tag{2.2}$$

This naturally translates into an iterative scheme, that alternate between the construction of a refined bound in the so-called minorization step and its subsequent maximization, leading to the acronym MM scheme. The descent property from equation (2.2) endows MM routines with remarkable numerical stability, as mild conditions for the target function ensure convergence of corresponding iterations (Lange, 2016).

EM algorithms as special instances of MM

MM schemes have been successfully employed for the optimization of a large variety of target functions $\mathcal{F}(\boldsymbol{\beta})$, which are typically minorized by purely mathematical arguments. However, when the target function corresponds to the log-likelihood $\ell(\boldsymbol{\beta}) = \log p(\mathbf{y} \mid \boldsymbol{\beta})$ of suitable statistical constructions, there is a large subclass of MM routines, known as expectation-maximization algorithms, whose distinguishing characteristic is to hinge on missing data representations (McLachlan & Krishnan, 1996). The core of the EM rationale consists in addressing the original likelihood $p(\mathbf{y} \mid \boldsymbol{\beta})$ as corresponding to an incomplete-data problem, hindered by the unavailability of a set of missing variables $\mathbf{z} \in \mathcal{Z}$ from some suitable space $\mathcal{Z}$. In particular, the interest lies in identifying a complete data space such that the maximization of $p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\beta})$ over $\boldsymbol{\beta}$ would result considerably more tractable compared to that of the starting problem. Therefore, as the augmented variables $\mathbf{z}$ remain partially or entirely unobserved, the complete log-likelihood is conveniently replaced by its conditional expectation given the observed data and the current value of the parameter

$$\mathcal{Q}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = \mathbb{E}_{p(\mathbf{z} \mid \boldsymbol{\beta}^{(t)}, \mathbf{y})} \big[\log p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\beta}) \big], \tag{2.3}$$

that is maximized to obtain an updated estimate $\boldsymbol{\beta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\beta}} \mathcal{Q}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$. As before, this scheme is iterated alternating between the update of the surrogate target in the so-called E-step and its optimization in the M-step, where the convergence of such procedure is guaranteed by the same monotonicity property presented above. Indeed, the EM procedure can be reformulated so as to give a proper minorizer as in equation (2.1)

$$\begin{aligned} \mathcal{G}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) &= \ell(\boldsymbol{\beta}^{(t)}) + \mathcal{Q}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) - \mathcal{Q}(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{\beta}^{(t)}) \\ &= \mathbb{E}_{p(\mathbf{z} \mid \boldsymbol{\beta}^{(t)}, \mathbf{y})} \left[\log \frac{p(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\beta})}{p(\mathbf{z} \mid \boldsymbol{\beta}^{(t)}, \mathbf{y})} \right], \end{aligned} \tag{2.4}$$

which allows to formally embed EM schemes within the broader framework of the MM rationale. In fact, the discrepancy between the log-likelihood in any two locations can be decomposed as the difference between the corresponding $\mathcal{Q}$ functions and a suitable Kullback-Leiber divergence (Kullback & Leibler, 1951), which is non-negative by definition

$$\ell(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}^{(t)}) = \mathcal{Q}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) - \mathcal{Q}(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{\beta}^{(t)}) + \mathrm{KL}\big[p(\mathbf{z} \mid \boldsymbol{\beta}^{(t)}, \mathbf{y}) \big\| p(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y}) \big].$$

Besides such minorization perspective, the explicit use of missing data representation endows EM schemes with several additional desirable properties. Above all, it provides a clear probabilistic interpretation of the procedure, that allows to draw direct connections with analogous Gibbs Sampling schemes (Gelfand, 2000) and variational Bayes approximation (Ormerod & Wand, 2010; Blei et al., 2017). At the same, it often facilitates the analytical study of its theoretical properties, particularly when the complete-data log-likelihood be-

longs to the regular exponential family of distributions (McLachlan & Krishnan, 1996).

2.2.1 MM via quadratic tangent bounds

The tractability of the surrogate problems plays a pivotal role in determining the practical efficiency of the corresponding MM scheme. For this reason, quadratic bounds of the form

$$\mathcal{G}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = \mathcal{F}(\boldsymbol{\beta}^{(t)}) + \nabla\mathcal{F}(\boldsymbol{\beta}^{(t)})^{\intercal}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^{\intercal}\mathbf{G}(\boldsymbol{\beta}^{(t)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) \tag{2.5}$$

have received considerable attention in several contexts (Wu & Lange, 2010), both within MM and EM formulations. Notice that the degrees of freedom in choosing the above surrogates in equation (2.5) concentrate solely on the curvature $\mathbf{G}(\boldsymbol{\beta}^{(t)}) = \nabla^2_{\boldsymbol{\beta}}\mathcal{G}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$, as the values of the constant and linear terms are imposed by the constraints in equation (2.1). The optimization of the above quadratic minorizer readily results in the updated vector

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\mathcal{M}}(\boldsymbol{\beta}^{(t)}) = \boldsymbol{\beta}^{(t)} - \big(\mathbf{G}(\boldsymbol{\beta}^{(t)})\big)^{-1}\nabla\mathcal{F}(\boldsymbol{\beta}^{(t)}), \tag{2.6}$$

where it is worth highlighting the parallel between the updates in equation (2.6) and the analogous ones corresponding to Newton-Raphson's method, for which the inverse of the exact Hessian $(\nabla^2\mathcal{F}(\boldsymbol{\beta}^{(t)}))^{-1}$ would take the place of $(\mathbf{G}(\boldsymbol{\beta}^{(t)}))^{-1}$. As already mentioned, the Newton-Raphson updates can be prone to oscillating or diverging behaviors, since they do not benefit from the monotonicity property that characterizes MM schemes. While equations (2.2) and (2.6) ensure respectively convergence of the procedure and tractability of the corresponding updates, the empirical performances of a specific MM scheme will be determined by the number of iterations needed for the difference $\mathcal{F}(\boldsymbol{\beta}^{(t+1)}) - \mathcal{F}(\boldsymbol{\beta}^{(t)})$ to become smaller than a fixed threshold. Accordingly, the convergence rate of MM schemes based on smooth surrogates is typically assessed by studying the spectral radius $\mathcal{R}(\boldsymbol{\beta})$ of the Jacobian $\boldsymbol{\mathcal{J}}(\beta)$ of the transformation map $\boldsymbol{\mathcal{M}} : \Re^p \to \Re^p$, describing the MM updates (Lange, 2016). If the objective $\mathcal{F}$ and the surrogate $\mathcal{G}$ are twice differentiable at a local maximum $\boldsymbol{\beta}^*$ and the corresponding Hessians are positive definite, it can be shown that

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\beta}^*) = \mathbf{I}_p - \big(\mathbf{G}(\boldsymbol{\beta}^*)\big)^{-1}\nabla^2\mathcal{F}(\boldsymbol{\beta}^*), \tag{2.7}$$

which gives $\mathcal{R}(\boldsymbol{\beta}^*) = 1 - \min_{\boldsymbol{\beta}\neq\mathbf{0}}\{\boldsymbol{\beta}^{\intercal}\nabla^2\mathcal{F}(\boldsymbol{\beta}^*)\boldsymbol{\beta} \,/\, \boldsymbol{\beta}^{\intercal}\mathbf{G}(\boldsymbol{\beta}^*)\boldsymbol{\beta}\}$ (Wu & Lange, 2010). This allows to get a clear intuition for the impact of the specific minorization considered, as the rate of convergence in proximity of a given equilibrium point will be roughly determined by how closely the curvature of the minorizer approximates that of the target. Accordingly, the tighter the approximation given by a particular surrogate is, the faster the convergence of the resulting MM scheme is expected to be.

2.2.2 Tangent bounds for logistic log-likelihoods

As mentioned above, the target function $\mathcal{F}(\boldsymbol{\beta})$ we aim at maximizing is the logistic log-likelihood $\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta})$, for which we have

$$\begin{aligned} \ell_i(\boldsymbol{\beta}) &= \log p(y_i \mid \boldsymbol{\beta}, \mathbf{x}_i) = y_i \mathbf{x}_i^\intercal \boldsymbol{\beta} - \log(1 + e^{\mathbf{x}_i^\intercal \boldsymbol{\beta}}) \\ \nabla \ell_i(\boldsymbol{\beta}) &= (y_i - \pi_i)\mathbf{x}_i \\ \nabla^2 \ell_i(\boldsymbol{\beta}) &= -\pi_i(1 - \pi_i)\mathbf{x}_i \mathbf{x}_i^\intercal, \end{aligned}$$

where $\pi_i = \pi(\mathbf{x}_i^\intercal \boldsymbol{\beta})$, while from now on we omit the covariates from the conditioning set for ease of notation. Notice that each term $\ell_i(\boldsymbol{\beta})$ depends on the parameters $\boldsymbol{\beta}$ only via the inner product with the corresponding covariates $\mathbf{x}_i^\intercal \boldsymbol{\beta}$, so that the gradient of the target function takes the form $\nabla \ell(\boldsymbol{\beta}) = \nabla \mathcal{F}(\boldsymbol{\beta}) = \mathbf{X}(\mathbf{y} - \boldsymbol{\pi})$, with $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)^\intercal$ and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\intercal$. At the same time, the element-wise minorization scheme entails a composite structure for the quadratic contribution of the bound $\mathbf{G}(\boldsymbol{\beta}^{(t)}) = -\mathbf{X}^\intercal \mathbf{W}^{(t)} \mathbf{X}$, where $\mathbf{W}^{(t)} = \mathbf{W}(\boldsymbol{\beta}^{(t)})$ is a positive definite $n \times n$ diagonal matrix. Indeed, taking advantage of the additive structure of the log-likelihood, the construction of a surrogate for the overall target is more often carried out by providing a tangent bound for the log-likelihood contribution of each statistical unit, since the minorization relation in equation (2.1) is closed respect to sum and non-negative product, other than limits and composition with increasing functions (Wu & Lange, 2010). As such, the overall tangent bound will be decomposed as

$$\mathcal{G}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = \sum_{i=1}^{n} g_i(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}).$$

MM via uniform quadratic bounds

Albeit the construction of a proper tangent bound is typically highly specific to each particular optimization problem, there is a set of mathematical tools frequently employed in the MM literature (Wu & Lange, 2010). Among others, uniform quadratic minorizers are often the first option considered when dealing with twice-differentiable concave target functions $\mathcal{F}$ with bounded curvature. In fact, if it exists a positive definite constant matrix $\mathbf{B}$ such that the difference $\mathbf{B} - \nabla^2 \mathcal{F}(\boldsymbol{\beta})$ is non-negative definite for every $\boldsymbol{\beta} \in \Re^p$, then a valid quadratic bound is simply obtained by fixing $\mathbf{G}(\boldsymbol{\beta}^{(t)}) = -\mathbf{B}$ in equation (2.5). In the case of logistic regression, such uniform minorization arises by recognizing that $\pi_i(1 - \pi_i) \in (0, 1/4]$, as firstly exploited in the seminal contribution by Böhning & Lindsay (1988). Accordingly, the quadratic function

$$g_{\text{BL},i}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = \ell_i(\boldsymbol{\beta}^{(t)}) + (y_i - \pi_i^{(t)})(\mathbf{x}_i^\intercal \boldsymbol{\beta} - \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)}) - \frac{1}{2}\frac{1}{4}(\mathbf{x}_i^\intercal \boldsymbol{\beta} - \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})^2$$

is a valid minorizer for the i-th log-likelihood term $\ell_i(\boldsymbol{\beta})$ at $\boldsymbol{\beta}^{(t)}$, where $\pi_i^{(t)} = \pi(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})$. Equivalently, this leads to an overall quadratic surrogates $\mathcal{G}_{\text{BL}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = \sum_{i=1}^n g_{\text{BL},i}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ whose curvature reads $\mathbf{G}_{\text{BL}}(\boldsymbol{\beta}^{(t)}) = -\mathbf{X}^\intercal \mathbf{W}_{\text{BL}} \mathbf{X}$, where $\mathbf{W}_{\text{BL}} = 1/4\, \mathbf{I}_n$.

EM via Pólya-Gamma data-augmentation

The EM rationale has proven to be extremely effective even in a variety of situations where the incompleteness of the data is not necessarily self-evident, beyond classical scenarios manifestly involving partially-observed data. In the case of logistic regression, a number of latent variable representations has been proposed over the years (Holmes & Held, 2006; Frühwirth-Schnatter & Frühwirth, 2007; Frühwirth-Schnatter et al., 2009; Polson et al., 2012; Zens et al., 2020). Among others, Polson et al. (2012) constructed a powerful data augmentation strategy by exploiting a scale-mixture representation for logistic likelihoods, which translates into the introduction of a Pólya-Gamma latent variable $z_i \in (0, \infty)$ for each statistical unit $i = 1, \ldots, n$, such that $(z_i \mid \boldsymbol{\beta}) \overset{\text{ind}}{\sim} \text{PG}(1, \mathbf{x}_i^\intercal \boldsymbol{\beta})$. Accordingly, the usual likelihood for the observed i-th binary observation $p(y_i \mid \boldsymbol{\beta})$ can be addressed as marginalization over z_i of a complete likelihood $p(y_i, z_i \mid \boldsymbol{\beta})$. Notably, the logarithm of the latter is a tractable quadratic function of the parameter $\boldsymbol{\beta}$, which largely facilitates both Frequentist and Bayesian inference. Indeed, Polson et al. (2012) exploited the restored conjugacy with the commonly-used Gaussian prior for the coefficients to build an efficient Gibbs sampling scheme with conjugate full-conditionals. Similarly, Durante & Rigon (2019) leveraged on the same hierarchical representation to construct both a formal mean-field variational approximation of the joint posterior $p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})$, with $\mathbf{z} = (z_1, \ldots, z_n)^\intercal$, other than a proper EM routine for maximum likelihood estimation. In particular, the lower bound on the i-th log-likelihood contribution exploited within the latter takes the form

$$
\begin{aligned}
g_{\text{PG},i}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) &= \mathbb{E}_{p(z_i \mid \boldsymbol{\beta}^{(t)})} \left[\log \frac{p(y_i, z_i \mid \boldsymbol{\beta})}{p(z_i \mid \boldsymbol{\beta}^{(t)})} \right] \\
&= \ell_i(\boldsymbol{\beta}^{(t)}) + (y_i - 1/2)\big(\mathbf{x}_i^\intercal \boldsymbol{\beta} - \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)}\big) - \frac{1}{2} w_{\text{PG}}(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})\big((\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 - (\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})^2\big)
\end{aligned}
$$

and $w_{\text{PG}}(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)}) = \tanh(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)}/2)/(2\, \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})$ arises as the expected value of a $\text{PG}(1, \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})$ random variable. As before, this translates into an overall quadratic surrogate $\mathcal{G}_{\text{PG}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = \sum_{i=1}^n g_{\text{PG},i}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ with curvature $\mathbf{G}_{\text{PG}}(\boldsymbol{\beta}^{(t)}) = -\mathbf{X}^\intercal \mathbf{W}_{\text{PG}}^{(t)} \mathbf{X}$, where now $\mathbf{W}_{\text{PG}}^{(t)} = \text{diag}(\{w_{\text{PG}}(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})\}_{i=1}^n)$.

As a matter of fact, the above quadratic minorizer was originally derived in the seminal contribution by Jaakkola & Jordan (2000). However, the authors did not appeal to any explicit missing data representation, leveraging instead solely on convexity arguments exploited in a more traditional-MM fashion, as detailed in Section 2.2.3. Their methodology has been successfully employed in the literature both for variational Bayes inference and ML estimation (Bishop & Svensén, 2003; Rasmussen & Williams, 2006; Lee et al., 2010; Ren

et al., 2011; Carbonetto & Stephens, 2012), although it lacked a clear probabilistic interpretation until the contribution by Durante & Rigon (2019). A first intuition on the relation with Pólya-Gamma data augmentation scheme appeared in the manuscript by Scott & Sun (2013), albeit the authors did not establish a full probabilistic connection, as the one settled by Durante & Rigon (2019).

Relative tightness and computational efficency

Durante & Rigon (2019) additionally carried out a direct comparison of the two iterative optimization schemes, exploiting respectively the quadratic bounds arising from the Pólya-Gamma data augmentation and the uniform one by Böhning & Lindsay (1988). Indeed, the authors showed that the former achieve a better asymptotic rate of convergence in proximity of any local maximum $\boldsymbol{\beta}^*$, and it thus is expected to lead to a faster convergence in practice. To do so, Durante & Rigon (2019) leveraged the proper EM formulation of the proposed scheme, by advocating the so-called missing information principle (McLachlan & Krishnan, 1996). This means that the Jacobian of the corresponding transformation matrix can be written as $\boldsymbol{\mathcal{J}}_{\text{PG}}(\boldsymbol{\beta}) = \mathbf{I}_p + \boldsymbol{\mathcal{I}}_c^{-1}(\boldsymbol{\beta}, y)\nabla^2\ell(\boldsymbol{\beta})$, where $\boldsymbol{\mathcal{I}}_c(\boldsymbol{\beta}, y)$ denotes the expectation, taken with respect to the augmented data, of the complete-data information matrix. This allowed the authors to simplify the comparison between the asymptotic rate of convergences of the two alternative optimization schemes. However, the same result can be seen also as a direct consequence of the overall quality of the respective resulting approximation for the target function. Indeed, on one hand equation (2.7) suggests that, for general MM schemes, the convergence rate reflects how closely the curvature of the minorizer approximates that of the target in proximity of a given local extremum. At the same time, since the curvature represents the only degree of freedom in constructing tangent quadratic minorizers, the relative quality of the two approximations concerns the bounds in their entirety. In fact, noticing that $w_{\text{PG}}(\mathbf{x}_i^\intercal\boldsymbol{\beta}^{(t)}) \in (0, 1/4]$ for any $\boldsymbol{\beta}^{(t)}$ and $\mathbf{x}_i$, it follows that the difference between the Hessian matrices

$$\mathbf{G}_{\text{PG}}(\boldsymbol{\beta}^{(t)}) - \mathbf{G}_{\text{BL}}(\boldsymbol{\beta}^{(t)}) = -\mathbf{X}^\intercal\mathbf{W}_{\text{PG}}^{(t)}\mathbf{X} + \mathbf{X}^\intercal\mathbf{W}_{\text{BL}}\mathbf{X}$$

is non-negative definite for every $\boldsymbol{\beta}^{(t)} \in \Re^p$. This means that the BL bound acts in turn as a uniform quadratic minorizer for the PG bound as well, meaning that

$$\ell(\boldsymbol{\beta}) \geq \mathcal{G}_{\text{PG}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) \geq \mathcal{G}_{\text{BL}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) \qquad \forall \boldsymbol{\beta} \in \Re^p.$$

Furthermore, the same argument can be further extended to prove that the PG bound is optimal among quadratic minorizers for the logistic log-likelihood, as detailed in Section 2.2.3.

Nevertheless, a thorough comparison of the alternative minorization schemes needs to take into account not only the relative speed of convergence, but also the computational

efficiency of the corresponding updates from equation (2.6). Indeed, the major bottleneck in performing such updates is typically given by a $p \times p$ matrix inversion, $(\mathbf{X}^\intercal \mathbf{W}_{\text{BL}} \mathbf{X})^{-1}$ and $(\mathbf{X}^\intercal \mathbf{W}_{\text{PG}}^{(t)} \mathbf{X})^{-1}$ respectively, which becomes prohibitive in high-dimensional scenarios. However, the uniform bound offers the possibility to perform such expensive operation only once as a pre-computation of the optimization procedure, greatly mitigating the cost of each update. On the contrary, the higher flexibility of the PG bound comes at the cost of requiring such costly matrix inversion at every update. Accordingly, the relative efficiency of the two minorization schemes will be determined by a trade-off between the associated convergence rate and the cost per iteration, largely influenced by the dimensionality of the specific regression problem considered. Nonetheless, empirical evidence suggests that the BL bound may suffer from poor convergence rates in the same scenarios where Newton-Raphson's method results unreliable, undermining the very argument motivating the use of a MM scheme.

2.2.3 Optimality of the PG bound among quadratic tangent minorizers

In this Section we further elaborate on the comparison between alternative quadratic minorizers for the logistic log-likelihood, clarifying the optimality of the PG bound within such a family. Beforehand, it is insightful to provide an alternative derivation of the PG bound, along the lines of its original formulation by Jaakkola & Jordan (2000). The authors proceed by first decomposing and symmetrizing each likelihood contribution $\ell_i(\boldsymbol{\beta})$, viewed as a function of the linear predictor $r_i = \mathbf{x}_i^\intercal \boldsymbol{\beta}$

$$\begin{aligned} \ell_i(r_i) &= (y_i - 0.5) r_i + h(r_i) \\ h(r_i) &= -\log(e^{r_i/2} + e^{-r_i/2}), \end{aligned} \tag{2.8}$$

where the source of intractability is condensed within the even function $h : \Re \to \Re^-$. Dropping the index i of the statistical unit for ease of notation, $h(r)$ is then reparametrized as a function of the squared linear predictor $\rho = r^2$, and lower bounded via its tangent surface at a given location $\varphi = \zeta^2$

$$h(\rho) \geq h(\varphi) + \frac{\partial h}{\partial \rho}(\varphi)(\rho - \varphi), \tag{2.9}$$

by the convexity of $h(\rho)$, while the same relation holds true even transforming back the problem in the original space

$$\begin{aligned} h(r) \geq h_{\text{PG}}(r \mid \zeta) &= h(\zeta) + \frac{\partial h}{\partial (r^2)}(\zeta^2)(r^2 - \zeta^2) = h(\zeta) - \frac{\tanh(\zeta/2)}{4\zeta}(r^2 - \zeta^2) \\ &= h(\zeta) + \frac{\partial h}{\partial r}(\zeta)(r - \zeta) - \frac{1}{2}\frac{\tanh(\zeta/2)}{2\zeta}(r - \zeta)^2. \end{aligned}$$

Recalling that $w_{\text{PG}}(\zeta) = \tanh(\zeta/2)/(2\zeta)$, it is immediate that the lower bound above coincides with PG minorizer $g_{\text{PG},i}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$, which is recovered by substituting r and ζ with $x_i^\intercal \boldsymbol{\beta}$ and $x_i^\intercal \boldsymbol{\beta}^{(t)}$ respectively. Conversely, the BL bound corresponds to the uniform minorization

$$h(r) \geq h_{\text{BL}}(r \mid \zeta) = h(\zeta) + \frac{\partial h}{\partial r}(\zeta)(r-\zeta) - \frac{1}{2}\frac{1}{4}(r-\zeta)^2.$$

While we already emphasized that $h_{\text{PG}}(r \mid \zeta) \geq h_{\text{BL}}(r \mid \zeta)$ for any $r \in \Re$ given any $\zeta \in \Re$, the same property can be further generalized to the comparison with any alternative quadratic tangent bound $h_{\text{Q}}(r \mid \zeta)$ within the family

$$\mathcal{H}_{\text{Q}}(\zeta) = \left\{ h_{\text{Q}}(r \mid \zeta) = h(\zeta) + \frac{\partial h}{\partial r}(\zeta)(r-\zeta) - \frac{1}{2} w_{\text{Q}}(\zeta)(r-\zeta)^2 \;\middle|\; \begin{array}{ll} h(\zeta) = h_{\text{Q}}(\zeta \mid \zeta) & \\ h(r) \geq h_{\text{Q}}(r \mid \zeta) & \forall r \end{array} \right\}.$$

Indeed, De Leeuw & Lange (2009) exploited the symmetry of the target function and of $h_{\text{PG}}(r \mid \zeta)$ to prove that the PG minorization provides the sharpest quadratic tangent bound that can be constructed for $h(r)$, meaning that

$$h_{\text{PG}}(r \mid \zeta) \geq h_{\text{Q}}(r \mid \zeta) \qquad \forall r \in \Re, \tag{2.10}$$

given any $\zeta \in \Re$ and $h_{\text{Q}}(r \mid \zeta) \in \mathcal{H}_{\text{Q}}(\zeta)$. The argument of De Leeuw & Lange (2009) can be rephrased by noticing that the specific condition $h_{\text{Q}}(\zeta \mid \zeta) = h(\zeta) = h(-\zeta) \geq h_{\text{Q}}(-\zeta \mid \zeta)$ directly translates into a bound on the curvature coefficient $w_{\text{Q}}(\zeta) \geq -\frac{\partial h}{\partial r}(\zeta)/\zeta = w_{\text{PG}}(\zeta)$, which is attained by the PG minorizer. While this is reflected in the symmetry of the latter, implying in particular that the PG bound is tangent to $h(r)$ both in ζ and $-\zeta$, the comparison with any alternative valid curvature ensures the relation in equation (2.10), as illustrated in Figure 2.1.

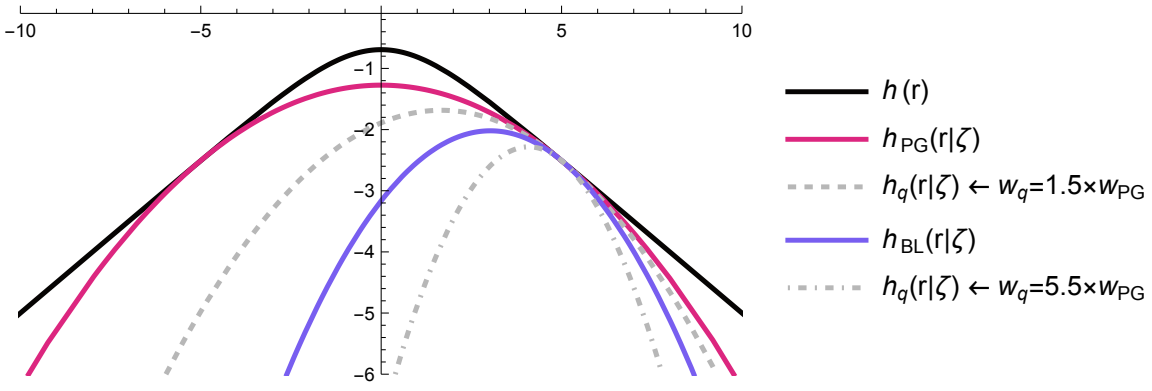

Figure 2.1: Comparison between different quadratic bounds for $h(r)$, tangent to the latter in $\zeta = 5$. $h_{\text{PG}}(r \mid \zeta)$ arises from the PG DA, $h_{\text{BL}}(r \mid \zeta)$ from a uniform minorization of the curvature, while the two quadratic lower bounds $h_{\text{Q}}(r \mid \zeta)$ corresponds respectively to $w_{\text{Q}}(\zeta) = 1.5 \cdot w_{\text{PG}}(\zeta)$ and $w_{\text{Q}}(\zeta) = 5.5 \cdot w_{\text{PG}}(\zeta)$.

Since the interest in deriving a minorizer for the log-likelihood as a function of the parameter $\boldsymbol{\beta} \in \Re^p$, the above univariate findings can be exploited to state the following result

Lemma 2.1. *Let $\mathcal{G}_{\mathrm{Q}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ be any quadratic tangent minorizer in $\boldsymbol{\beta}^{(t)} \in \Re^p$ for the logistic log-likelihood $\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\beta}) = \sum_{i=1}^n \left(y_i \mathbf{x}_i^\intercal \boldsymbol{\beta} - \log(1 + e^{\mathbf{x}_i^\intercal \boldsymbol{\beta}})\right)$. Furthermore, let $\mathcal{G}_{\mathrm{Q}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ be separable as the sum of the contributions associated with each statistical unit, with each contribution depending on $\boldsymbol{\beta}$ only via the inner product with the corresponding linear predictor*

$$\mathcal{G}_{\mathrm{Q}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = \sum_{i=1}^n k_{\mathrm{Q},i}(\mathbf{x}_i^\intercal \boldsymbol{\beta} \mid \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)}),$$

for a suitable set of quadratic functions $\{k_{\mathrm{Q},i}\}_{i=1}^n$. Then it holds that

$$\ell(\boldsymbol{\beta}) \geq \mathcal{G}_{\mathrm{PG}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) \geq \mathcal{G}_{\mathrm{Q}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) \qquad \forall \boldsymbol{\beta} \in \Re^p,\ \forall \boldsymbol{\beta}^{(t)} \in \Re^p.$$

Proof. As highlighted before, the requirement that $\mathcal{G}_{\mathrm{Q}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ minorizes $\ell(\boldsymbol{\beta})$ at $\boldsymbol{\beta}^{(t)}$ constraints the first two terms of every quadratic bound

$$\mathcal{G}_{\mathrm{Q}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = \ell(\boldsymbol{\beta}^{(t)}) + \nabla \ell(\boldsymbol{\beta}^{(t)})^\intercal (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^\intercal \boldsymbol{\mathcal{A}}_{\mathrm{Q}}(\boldsymbol{\beta}^{(t)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}),$$

whereas separability ensures that the $p \times p$ positive definite matrix $\boldsymbol{\mathcal{A}}_{\mathrm{Q}}(\boldsymbol{\beta}^{(t)})$ is decomposed as $\boldsymbol{\mathcal{A}}_{\mathrm{Q}}(\boldsymbol{\beta}^{(t)}) = \mathbf{X}^\intercal \mathbf{W}_{\mathrm{Q}}(\boldsymbol{\beta}^{(t)}) \mathbf{X}$, where $\mathbf{W}_{\mathrm{Q}}(\boldsymbol{\beta}^{(t)}) = \mathrm{diag}(\{w_{\mathrm{Q},i}(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})\}_{i=1}^n)$ is an $n \times n$ diagonal matrix. Accordingly, the optimality of $\mathcal{G}_{\mathrm{PG}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ as a function of $\boldsymbol{\beta}$ arises directly as by contradiction, since the existence of $\boldsymbol{\beta}^{(t)}$, $\{w_{\mathrm{Q},i}(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})\}_{i=1}^n$ and $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(t)})$ such that $\ell(\tilde{\boldsymbol{\beta}}) \geq \mathcal{G}_{\mathrm{Q}}(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{\beta}^{(t)}) > \mathcal{G}_{\mathrm{PG}}(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{\beta}^{(t)})$ would negate the result in equation (2.10). □

2.3 Logistic regression under elastic net penalty

In penalized estimation contexts, the log-likelihood $\ell(\boldsymbol{\beta})$ is coupled with a regularization term $\mathcal{P}(\boldsymbol{\beta})$, with the goal of enhancing both prediction accuracy and interpretability. This is of particular interest in high dimensional regimes, where the number of covariates p is considerably greater than the number of observations n. In this Section, we focus on logistic regression under one of the most popular regularization methods, namely the elastic net by Zou & Hastie (2005), which gives a compromise between lasso and ridge regularizers (Hoerl & Kennard, 1981; Tibshirani, 1996)

$$\mathcal{P}(\boldsymbol{\beta}) = \lambda\Big((1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1\Big). \tag{2.11}$$

On one hand, it inherits the good predictive power of ridge regression. At the same time, it behaves similarly to the lasso in discarding ineffective predictors. The parameter $\lambda \in \Re^+$

in equation (2.11) determines the overall strength of the regularization, while $\alpha \in (0,1)$ regulates the relative magnitude of the ℓ_1 and ℓ_2-norm contributions. The resulting penalized optimization problem is often formulated as the minimization $\min_{\boldsymbol{\beta}} \big\{ -\ell(\boldsymbol{\beta}) + \mathcal{P}(\boldsymbol{\beta}) \big\}$. Nonetheless, we here choose to phrase it rather as a maximization problem by a simple sign change, for coherence with the setting of the previous Sections.

Indeed, the MM and EM schemes presented above are readily extended to such penalized ML estimation problems, thanks to the aforementioned closure of the minorization operation with respect to addition and non-negative product. Accordingly, the minorization step still proceeds by providing a valid tangent bound for each log-likelihood term $\ell_i(\boldsymbol{\beta})$, exploited to construct an overall minorizer $\mathcal{G}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ for $\mathcal{F}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \mathcal{P}(\boldsymbol{\beta})$. Nonetheless, the maximization step is now hindered by the further intractability introduced by ℓ_1-norm in the regularization, which prevents closed-form solutions for the optimization in $\Re^p$ even in combination with simple quadratic bounds for the log-likelihood terms. In the latter case, the overall surrogate takes the form

$$\mathcal{G}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = -\frac{1}{2} \sum_{i=1}^{n} w_i^{(t)} \big(\tau_i^{(t)} - \mathbf{x}_i^{\intercal} \boldsymbol{\beta} \big)^2 - \lambda(1-\alpha) \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 - \lambda \alpha \|\boldsymbol{\beta}\|_1 + \text{const}, \tag{2.12}$$

where the form for the weights $w_i^{(t)} = w_i(\mathbf{x}_i^{\intercal} \boldsymbol{\beta}^{(t)})$ and effective residuals $\tau_i^{(t)} = \tau_i(\mathbf{x}_i^{\intercal} \boldsymbol{\beta}^{(t)})$ will be determined by the specific minorization procedure considered. The resulting optimization problem can be formally regarded as an instance of quadratic programming, for which several sophisticated methods have been developed over the years (Nesterov & Nemirovskii, 1994). However, such routines often turn out to be computationally sub-optimal compared to other more specific techniques, tailored for the lasso regularization. Indeed, while some contributions in the literature considered the possibility of deriving an additional quadratic bound for the penalty term (Wu & Lange, 2010), most popular approaches rely on coordinate-wise optimization schemes (Friedman et al., 2007, 2010), which amount to optimizing sequentially the surrogate target one coordinate β_j at a time, while keeping fixed the remaining $\beta_{j'}$, for every $j' \neq j$. This create an inner sequence of iterates $\{\boldsymbol{\beta}^{(t,s)}\}_{s \geq 0}$ starting from the current tangency location $\boldsymbol{\beta}^{(t,0)} = \boldsymbol{\beta}^{(t)}$, that are updated by setting $\beta_{j'}^{(t,s+1)} = \beta_{j'}^{(t,s)}$, for all $j' \neq j$ and

$$\beta_j^{(t,s+1)} = \text{argmax}_{\beta_j} \Big\{ \mathcal{G}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) \;\; \big| \; \beta_{j'} = \beta_{j'}^{(t,s)} \; \forall j' \neq j \Big\},$$

where $j = s \pmod p$ for the so-called cyclic version of coordinate descent, while $\boldsymbol{\beta}^{(t+1)}$ is eventually set to the final value of such inner cycle, once convergence has been reached.

In particular, the EM formulation of the PG minorization allows the embedding of the resulting coordinate-wise optimization scheme within the broader framework of expectation-conditional-maximization (ECM) algorithms. These are extensions of EM schemes, introduced by Meng & Rubin (1993) to deal with situations where even the complete-data ML

estimation remains rather complicated, not necessarily because of the combination with a non-smooth regularization. Accordingly, every M-step is replaced with a sequence of simpler conditional-maximization (CM) steps, each of which maximizes the conditional expectation of the target function obtained in the last E-step subject to a set of adaptive constraints, that are expressed as the conditioning on some function of the parameters. The inner cycle of CM-steps is typically run until convergence as in the formulation above, although it is also possible to fix a priori a given number of CM-steps per each E-step.

2.3.1 Cyclic coordinate-wise optimization

The inner sequence arising from such coordinate-wise optimization scheme is guaranteed to converge to the global optimum $\boldsymbol{\beta}^{(t+1)} = \text{argmax}_{\boldsymbol{\beta}} \mathcal{G}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$, provided the non-smooth penalty contribution is a separable function with respect to the coordinates of $\boldsymbol{\beta}$

$$\mathcal{P}(\boldsymbol{\beta}) = \textstyle\sum_{j=1}^{p} \mathcal{P}_j(\beta_j), \tag{2.13}$$

where $\mathcal{P}_j : \Re \to \Re$ are univariate convex non-differentiable functions (Hastie et al., 2015, Section 5.4.1). As the above criterion is clearly satisfied by elastic net penalized logistic regression, we can focus on one-dimensional surrogates of the kind

$$\begin{aligned} \mathcal{G}(\beta_j; \boldsymbol{\beta}^{(t,s)} \mid \boldsymbol{\beta}^{(t)}) = -\frac{1}{2} \sum_{i=1}^{n} w_i^{(t)} \cdot \left(x_{ij}\beta_j - \left(\tau_i^{(t)} + x_{ij}\beta_j^{(t,s)} - \mathbf{x}_i^{\intercal} \boldsymbol{\beta}^{(t,s)} \right) \right)^2 \\ - \lambda(1-\alpha)\frac{1}{2}\beta_j^2 - \lambda\alpha|\beta_j| + \text{const}, \end{aligned} \tag{2.14}$$

for each $j = 1, \ldots, p$. This reformulation of the maximization problem has a twofold advantage. On one hand, it avoids expensive large matrix inversions, in contrast to what happens for the unpenalized updates of equation (2.6). At the same time, it restores the tractability of the MM scheme, since the maximization of equation (2.14) allows for an exact solution of the form

$$\beta_j^{(t,s+1)} = \frac{\mathcal{S}\left(\sum_{i=1}^{n} w_i^{(t)} x_{ij} \left(\tau_i^{(t)} + x_{ij}\beta_j^{(t,s)} - \mathbf{x}_i^{\intercal} \boldsymbol{\beta}^{(t,s)} \right), \lambda\alpha \right)}{\sum_{i=1}^{n} w_i^{(t)} x_{ij}^2 + \lambda(1-\alpha)}, \tag{2.15}$$

where $\mathcal{S}(r, \delta)$ is the so-called soft-thresholding operator (Hastie et al., 2015)

$$\mathcal{S}(r, \delta) = \text{sign}(r)(|r| - \delta)_+ = \begin{cases} r - \delta & \text{if } \; r > 0 \; \text{ and } \; \delta < |r| \\ 0 & \text{if } \; \delta \geq |r| \\ r + \delta & \text{if } \; r < 0 \; \text{ and } \; \delta < |r| \; . \end{cases} \tag{2.16}$$

It is worth mentioning that advanced computational routines typically complement the described coordinate-wise optimization scheme with a set of significant heuristics, with

the goal of enhancing empirical performances. For instance, the aforementioned `glmnet` package implements the so-called path-wise coordinate optimization framework by Friedman et al. (2007), which makes use of warm-start initialization and active set convergence (Friedman et al., 2010). The former means that the solution is computed progressively for a decreasing sequence of values of the regularization parameter, rather than for a single one, using the outcome of one optimization cycle as the starting value for the following. The latter signifies that the optimization routine concentrates only on a dynamic subset of the coordinates of $\boldsymbol{\beta}$, namely the non-zero ones, updating such set at the end of each maximization cycle. While these approaches can be extremely advantageous in practice, in this work we choose not to employ the above heuristics, with the purpose of better appreciating the relative efficiency of the routines based on different lower bounds. As such, we focus on the optimization problem for a single value of the regularization parameter, performing the maximization for every coordinate of parameter $\boldsymbol{\beta}$. Nonetheless, all the methods presented in this Chapter can be readily integrated within the same path-wise coordinate optimization scheme.

Uniform and PG quadratic bounds and relative performance

A practical implementation of the above MM coordinate-wise optimization scheme for penalized ML estimation in logistic regression requires solely the specification of the weights $\{w_i^{(t)}\}_{i=1}^n$ and effective residuals $\{\tau_i^{(t)}\}_{i=1}^n$. Simple algebraic calculations show that the uniform quadratic bound by Böhning & Lindsay (1988) corresponds to

$$\begin{cases} w_{\text{BL},i}^{(t)} & = 1/4 \\ \tau_{\text{BL},i}^{(t)} & = \big((y_i - \pi_i^{(t)}) + w_{\text{BL},i}^{(t)}\, \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)}\big)/w_{\text{BL},i}^{(t)} \,. \end{cases}$$

The use of the above coefficients above can often be explicitly selected in several state-of-the-art statistical software as an alternative to the Newton-Raphson's updates, that would correspond to $w_{\text{NR},i}^{(t)} = \pi_i^{(t)}(1-\pi_i^{(t)})$ and $\tau_{\text{NR},i}^{(t)} = \big((y_i-\pi_i^{(t)})+w_{\text{NR},i}^{(t)}\, \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)}\big)/w_{\text{NR},i}^{(t)}$. Conversely, the quadratic surrogate arising for the Pólya-Gamma data augmentation translates into weights and effective residuals given by

$$\begin{cases} w_{\text{PG},i}^{(t)} & = \tanh(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)}/2)/(2\, \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)}) \\ \tau_{\text{PG},i}^{(t)} & = (y_i - 1/2)/w_{\text{PG},i}^{(t)} \,. \end{cases}$$

We highlight here that the comparison between the MM routines arising from the two alternative quadratic bounds presents a substantial difference compared to the unpenalized case, arising from the use of the coordinate-wise optimization scheme. Indeed, the iterative solution of the univariate maximization problems via equation (2.15) does not involve anymore any large matrix inversion, which encompassed the only significant operational

difference between the two quadratic minorization procedures in the absence of regularization. As a consequence, this eliminates the trade-off between cost per update and speed of convergence associated with the PG and BL bounds described in the previous Section. In fact, in the coordinate-wise optimization framework, the relative convenience of a particular bound will be determined solely by the resulting finite-time convergence rate, which as before is intuitively connected to the tightness of the minorization employed. Indeed, Lemma 2.1 still implies that

$$\ell(\boldsymbol{\beta}) - \mathcal{P}(\boldsymbol{\beta}) \geq \mathcal{G}_{\text{PG}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) - \mathcal{P}(\boldsymbol{\beta}) \geq \mathcal{G}_{\text{Q}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) - \mathcal{P}(\boldsymbol{\beta}) \qquad \forall \boldsymbol{\beta} \in \Re^p$$

for all $\boldsymbol{\beta}^{(t)} \in \Re^p$ and all tangent quadratic minorizer $\mathcal{G}_{\text{Q}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$, which in particular includes the BL bound. Accordingly, the PG minorization is expected to be consistently superior to the use of any alternative quadratic tangent bound, as supported by the empirical evidence reported in Section 2.5. Despite this fact, the PG bound is surprisingly not implemented in many of the most used packages addressing logistic regression under elastic-net regularization, such as the `glmnet` R package mentioned previously, in contrast to the uniform minorization by Böhning & Lindsay (1988). In fact, our results suggest a systematic replacement of the latter with the PG bound in all coordinate-wise optimization algorithms for logistic regression, requiring only minor practical modification. A more formal assessment of the relative convenience of the two MM schemes would require a quantitative analysis of the associated finite time convergence rates. However, as noted recently in Klopfenstein et al. (2020), the non-differentiability of the ℓ_1-norm in the penalty significantly hinders an analytical study of the convergence rate, which remains object of active research (Saha & Tewari, 2013; Zhao et al., 2018). In fact, the majority of the results in the literature addressing this issue for ECM and EM schemes are limited to smooth target and surrogate functions (Meng & Rubin, 1993; Meng, 1994; McLachlan & Krishnan, 1996; Ma et al., 2000; Zhou & Zhang, 2012).

2.4 Beyond PG optimality: piece-wise quadratic minorization

The key feature driving the interest in quadratic tangent bounds for logistic log-likelihoods dwells in their tractability, which in turn translates into the computational efficiency of the associated MM schemes. Nonetheless, the comparison between alternative quadratic minorizations stimulates the question regarding the possibility to develop a tighter one, especially in light of the aforementioned connection between the speed of convergence of a given MM scheme and the sharpness of the underlying surrogate functions. Indeed, several contributions in the literature focused on improving over the PG bound by exploiting piece-wise quadratic minorizers (Khan et al., 2010; Marlin et al., 2011; Ermis & Bouchard, 2014). In particular, Marlin et al. (2011) proposed the use of fixed minimax-optimal piece-wise quadratic bounds among all possible piece-wise quadratic tangent bounds for logistic log-

likelihoods

$$h_{\text{PQ}}(r; R) = \sum_{s=1}^{R} (a_s r^2 + b_s r + c_s) \cdot \mathbb{1}(r \in [t_{s-1}, t_s)).$$

In doing so, they consider the number of disjoint intervals R composing the domain of the surrogate function to be a principal tunable parameter, which regulates a trade-off between the accuracy and the complexity of the resulting approximation. For an arbitrary number of pieces R, the piece-wise quadratic bound is then constructed by solving numerically a minimax optimization problem both on the locations identifying the interval's separation and on the local coefficients of the quadratic contributions

$$\min_{\{a_s, b_s, c_s, t_s\}} \max_{s=1,\dots,R} \max_{r \in [t_{s-1}, t_s)} \Big(h(r) - h_{\text{PQ}}(r; R) \Big)$$
$$\left| \begin{array}{ll} h(r) - \big(a_s r^2 + b_s r + c_s\big) \geq 0 & \forall s, \forall r \in [t_{s-1}, t_s) \\ t_s - t_{s-1} > 0 & \forall s = 1, \dots, R \\ a_s \leq 0 & \forall s = 1, \dots, R\,, \end{array} \right.$$

further imposing bounded discrepancy from the target in each of the R sets. The output of such numeric optimization is then exploited within a generalized EM algorithm to overcome the intractability of some logistic-Gaussian integrals, replacing the intractable logistic log-likelihood with the fixed piece-wise bound. This separates the construction of the fixed bound from the learning phase of the inferential procedure, as the former is intended as a pre-computed approximation of an analytically intractable component of the model, whose accuracy is controlled via the cardinality of the underlying partitioning of the domain space. Furthermore, the generality of such a class of approximating functions comes at cost of the unavailability of a tractable analytical formulation.

In the present Section, we construct a novel piece-wise quadratic tangent bound of the logistic log-likelihoods, that addresses both the aforementioned limitations of the contribution by Marlin et al. (2011). Assuming again a separable structure of the overall bound $\mathcal{G}_{\text{PLQ}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = \sum_{i=1}^{n} g_{\text{PLQ},i}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$, we complement each quadratic term with an additional piece-wise linear contribution, proportional to the ℓ_1-norm of the associated linear predictor $|\mathbf{x}_i^\intercal \boldsymbol{\beta}|$. Accordingly, we chose the acronym PLQ (piece-wise linear-quadratic) to denote the resulting surrogate, as to emphasize the source of the piece-wise behavior. Our approach differs from several perspectives from that of Marlin et al. (2011), as detailed later in this Section.

2.4.1 Novel piece-wise linear-quadratic bound

The proposed PLQ bound is more easily constructed by working in the same transformed space as in Jaakkola & Jordan (2000), namely focusing on the reparametrization of $h(r)$ as a function of the squared linear predictor $\rho = r^2$. Indeed, the minorization in equation (2.9) can be improved only by introducing some curvature in the right and side, for instance by

complementing it with a term proportional to $\sqrt{\rho}$

$$h(\rho) \geq h_{\text{PLQ}}(\rho \mid \varphi) = h(\varphi) - \frac{1}{2} w_{\text{PLQ}}(\varphi)(\rho - \varphi) - \nu_{\text{PLQ}}(\varphi)(\sqrt{\rho} - \sqrt{\varphi}).$$

While the tangency condition typical of MM scheme $\frac{\partial h}{\partial \rho}(\varphi) = \frac{\partial h_{\text{PLQ}}(\rho|\varphi)}{\partial \rho}(\varphi)$ imposes a constraint of the two coefficients $w_{\text{PLQ}}(\varphi)$ and $\nu_{\text{PLQ}}(\varphi)$, the remaining degree of freedom can be optimized over by imposing the constraint $h(0) = h_{\text{PLQ}}(0 \mid \varphi)$, which leads to

$$\begin{aligned} w_{\text{PLQ}}(\varphi) &= \frac{2}{\varphi}\left(h(\varphi) - h(0) - 2\,\varphi\,\frac{\partial h}{\partial \rho}(\varphi)\right) \\ \nu_{\text{PLQ}}(\varphi) &= -\frac{2}{\sqrt{\varphi}}\left(h(\varphi) - h(0) - \varphi\,\frac{\partial h}{\partial \rho}(\varphi)\right). \end{aligned}$$

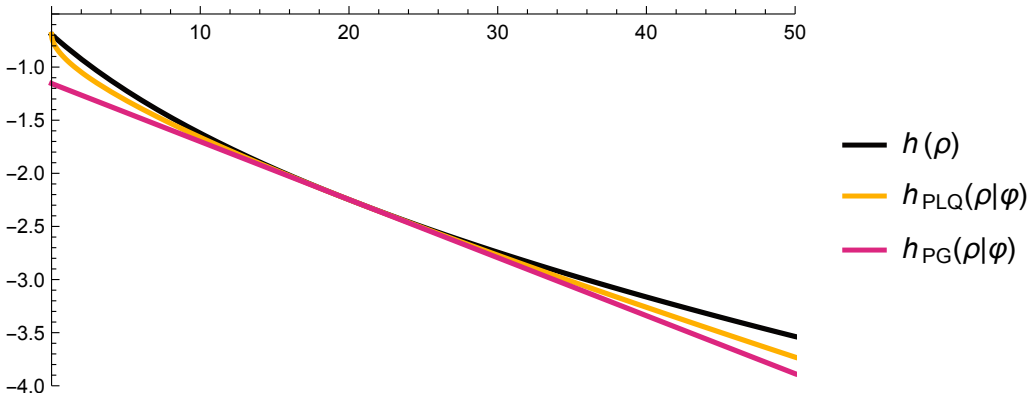

Figure 2.2: Comparison between the PLQ and the PG tangent bounds $h_{\text{PLQ}}(\rho \mid \varphi)$ and $h_{\text{PG}}(\rho \mid \varphi)$ for $h(\rho)$ as a function of the squared linear predictor $\rho = r^2$, with $\varphi = 20$.

While Figure 2.2 gives a visual representation of the resulting accuracy gain over the PG bound as a function of $\rho = r^2$, the minorization clearly remains valid even under the usual parametrization

$$h(r) \geq h_{\text{PLQ}}(r \mid \zeta) = h(\zeta) - \frac{1}{2} w_{\text{PLQ}}(\zeta)(r^2 - \zeta^2) - \nu_{\text{PLQ}}(\zeta)(|r| - |\zeta|), \tag{2.17}$$

where now

$$\begin{aligned} w_{\text{PLQ}}(\zeta) &= 2\,w_{\text{PG}}(\zeta) - 2\log\cosh\big(\zeta/2\big)/\zeta^2 \\ \nu_{\text{PLQ}}(\zeta) &= |\zeta|\big(w_{\text{PG}}(\zeta) - w_{\text{PLQ}}(\zeta)\big), \end{aligned} \tag{2.18}$$

while Figure 2.3 allows for an intuitive comparison between the relative accuracy of the different minorization schemes presented in this work.

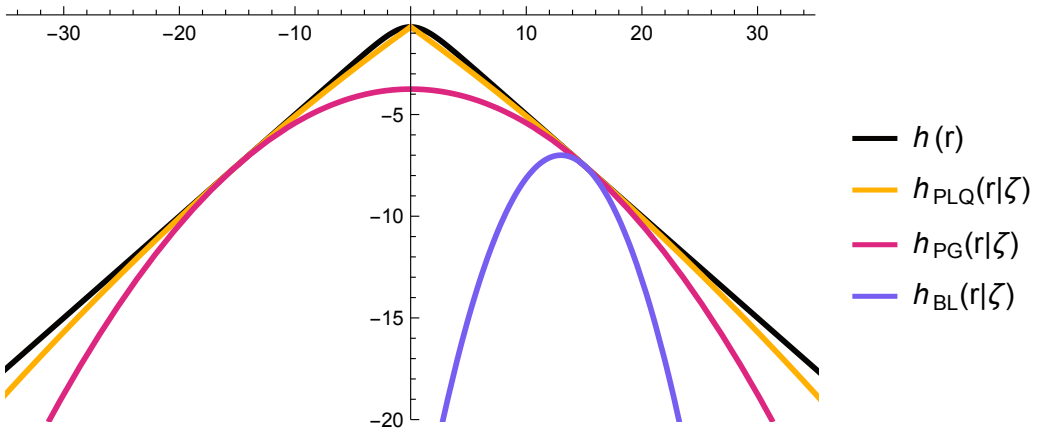

Figure 2.3: Comparison between the novel piece-wise linear-quadratic bound $h_{\text{PLQ}}(r \mid \zeta)$ and the usual quadratic minorizers $h_{\text{PG}}(r \mid \zeta)$ and $h_{\text{BL}}(r \mid \zeta)$ in the original space, with $\zeta = 20$. The three bounds coincide in the limit of $|\zeta|$ going to zero. Conversely, the larger is $|\zeta|$, the greater is the relative improvement in the approximation accuracy given by the PLQ over the PG one.

Comparison with PG bound and more involved piece-wise quadratic minorizations

Thanks to the above formulation, it is relatively easy to prove that the PLQ bound dominates over the PG bound

$$h(r) \geq h_{\text{PLQ}}(r \mid \zeta) \geq h_{\text{PG}}(r \mid \zeta),$$

by exploiting the majorization for the absolute value function $|r| \leq \frac{1}{2}\left(\frac{r^2}{|\zeta|} + |\zeta|\right)$, frequently employed in the MM literature (Hunter & Lange, 2004; Wu & Lange, 2010). Indeed, it can be easily verified that the usual PG bound is indeed recovered by substituting the piece-wise linear term in the PLQ minorizer with such quadratic tangent bound for the absolute value function. In particular, this is a direct consequence of the tangency condition for the PLQ bound, which gives

$$w_{\text{PLQ}}(\zeta) + \frac{1}{|\zeta|}\nu_{\text{PLQ}}(\zeta) = w_{\text{PG}}(\zeta)\ . \tag{2.19}$$

At the same time, proving a general optimality result, analogous to that of equation (2.10), appears to be more subtle in the case of the PLQ bound, as the piece-wise behavior introduces a significant degree of flexibility, already with only two quadratic branches. However, if we restrict our attention to the class of two-fold piece-wise quadratic minorizers for which the non-smooth behavior can be expressed as arising from the absolute value function

$$\mathcal{H}_{\text{S}}(\zeta) = \left\{ \begin{array}{c} h_{\text{S}}(r \mid \zeta) = h(\zeta) + a_{\text{S}}(\zeta)(r-\zeta) - \frac{1}{2}w_{\text{S}}(\zeta)(r^2-\zeta^2) - \nu_{\text{S}}(\zeta)(|r|-|\zeta|) \\ \text{s.t.} \quad h(\zeta) = h_{\text{S}}(\zeta \mid \zeta) \quad \text{and} \quad h(r) \geq h_{\text{S}}(r \mid \zeta) \quad \forall r \end{array} \right\}.$$

then we are able to prove that the PLQ bound dominates over all such minorizers

$$h_{\text{PLQ}}(r \mid \zeta) \geq h_{\text{S}}(r \mid \zeta) \qquad \forall r \in \Re, \tag{2.20}$$

given any $\zeta \in \Re \setminus \{0\}$ and $h_{\text{S}}(r \mid \zeta) \in \mathcal{H}_{\text{S}}(\zeta)$, as detailed in Appendix A.2.1.

Albeit implicitly included in the general family of piece-wise quadratic tangent bounds for the logistic log-likelihood, the methodology we propose differs in several aspects from that of Marlin et al. (2011). Above all, we provide an explicit analytical formulation of the minorizer, which entails only a single splitting point for the domain of each likelihood contribution at the origin, and restrict the degree of freedom by imposing the same curvature on both the resulting quadratic branches. In doing so, we avoid imposing bounded discrepancy from the target, since the increased flexibility of the PLQ bound already provides a substantial accuracy gain over the minorization schemes by Böhning & Lindsay (1988) and Jaakkola & Jordan (2000). Finally, the variables $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)^\intercal$, which parameterize the bound and its coefficients, are learned adaptively as part of the inferential procedure, instead of being pre-determined via data-agnostic numerical optimization.

Interpretation and connection with generalized lasso

The explicit formulation in terms of the ℓ_1-norms of the linear predictors further allows for a two-fold interpretation of the proposed minorizers. Indeed, the MM routines exploiting quadratic surrogates can be interpreted as tackling logistic regression by solving an adaptive sequence of re-weighted least-squares approximate problems. Conversely, the combination of the PLQ bound for $h(r)$ with the exact linear term from equation (2.8) results in the overall tangent minorizer

$$\begin{aligned} \mathcal{G}_{\text{PLQ}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = &\sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}^{(t)}) + \sum_{i=1}^{n} (y_i - 1/2)\Big(\mathbf{x}_i^\intercal \boldsymbol{\beta} - \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)}\Big) \\ &- \frac{1}{2} \sum_{i=1}^{n} w_{\text{PLQ}}(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})\Big((\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 - (\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})^2\Big) \\ &- \sum_{i=1}^{n} \nu_{\text{PLQ}}(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})\Big(|\mathbf{x}_i^\intercal \boldsymbol{\beta}| - |x_i^\intercal \beta^{(t)}|\Big). \end{aligned} \tag{2.21}$$

Alternatively, the above surrogate can be rewritten as

$$\mathcal{G}_{\text{PLQ}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = -n \log 2 + (\mathbf{y} - 0.5 \cdot \mathbf{1}_n)^\intercal \mathbf{X}\boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^\intercal \mathbf{X}^\intercal \mathbf{W}_{\text{PLQ}}^{(t)} \mathbf{X}\boldsymbol{\beta} - \|N_{\text{PLQ}}^{(t)} \mathbf{X}\boldsymbol{\beta}\|_1, \tag{2.22}$$

where $\mathbf{W}_{\text{PLQ}}^{(t)} = \text{diag}(\{w_{\text{PLQ}}(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})\}_{i=1}^n)$ and $\mathbf{N}_{\text{PLQ}}^{(t)} = \text{diag}(\{\nu_{\text{PLQ}}(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)})\}_{i=1}^n)$. This suggests that the essence of logistic regression is better grasped by an approximating sequence of adaptive combinations of re-weighted least-squares and least-absolute-deviations prob-

lems, which can be equivalently addressed as least-squares regressions under an adaptive generalized lasso penalty (Tibshirani & Taylor, 2011; Hastie et al., 2015; Arnold & Tibshirani, 2016; Ali & Tibshirani, 2019). The latter is an extension of the celebrated lasso regularization, which penalizes the regression via the ℓ_1-norm of a suitable linear transformation $\mathbf{D}\boldsymbol{\beta}$ of the parameters, instead of the usual penalization $\|\boldsymbol{\beta}\|_1$. The scope of such penalties is to enforce certain structural constraints on the coefficients, rather than pure sparsity, as it happens for instance in the case of the fused lasso (Tibshirani et al., 2005) and trend filtering (Kim et al., 2009). Indeed, the generalized lasso regularization induced by the proposed minorization scheme at the t-th step corresponds to the transformation matrix $\mathbf{D} = \mathbf{N}^{(t)}_{\text{PLQ}}\mathbf{X}$, which essentially enforces a penalization on values of $\boldsymbol{\beta}$ resulting in large values of the linear predictors $\mathbf{X}\boldsymbol{\beta}$, further strengthened by the monotonicity of the multiplicative coefficients $\nu_{\text{PLQ}}(\zeta)$ with respect to $|\zeta|$.

2.4.2 Bound optimization and MM algorithm under elastic net penalty

The price to be paid for the higher flexibility of the PLQ bound is the unavailability of a closed-form solution for the joint maximization steps on $\Re^p$ in the resulting MM algorithm, even for the unconstrained case. In fact, the nature of the resulting surrogate formally places the problem within the broader framework of the optimization of piece-wise quadratic functions with linear constraints (Lucet et al., 2009; Cui et al., 2020). However, it is computationally advantageous to leverage on the specific structure of the problem considered, similarly to what we have seen before in Section 2.3. Indeed, it can be shown that the resulting coordinate-wise optimization problems still admit exact solutions, which in particular suggest the use of PLQ in combination with ℓ_1-penalties, as they already dictate the use of coordinate descent schemes. In the latter case, the one-dimensional regularized surrogate would take the form

$$\begin{aligned}\mathcal{G}_{\text{PLQ}}(\beta_j; \boldsymbol{\beta}^{(t,s)} \mid \boldsymbol{\beta}^{(t)}) = &-\frac{1}{2}\sum_{i=1}^{n} w^{(t)}_{\text{PLQ},i} \cdot \left(x_{ij}\beta_j - \left(\tau^{(t)}_{\text{PLQ},i} + x_{ij}\beta_j^{(t,s)} - \mathbf{x}_i^\intercal\boldsymbol{\beta}^{(t,s)}\right)\right)^2 \\ &- \sum_{i=1}^{n} \nu^{(t)}_{\text{PLQ},i} \cdot \left|x_{ij}\beta_j - \left(x_{ij}\beta_j^{(t,s)} - \mathbf{x}_i^\intercal\boldsymbol{\beta}^{(t,s)}\right)\right| \\ &- \lambda(1-\alpha)\frac{1}{2}\beta_j^2 - \lambda\alpha|\beta_j| + \text{const}\,,\end{aligned}$$

with $w^{(t)}_{\text{PLQ},i} = w_{\text{PLQ}}(\mathbf{x}_i^\intercal\boldsymbol{\beta}^{(t)})$, $\nu^{(t)}_{\text{PLQ},i} = \nu_{\text{PLQ}}(\mathbf{x}_i^\intercal\boldsymbol{\beta}^{(t)})$ and $\tau^{(t)}_{\text{PLQ},i} = (y_i - 1/2)/w^{(t)}_{\text{PLQ},i}$, while the solution to the maximization problem $\beta_j^{(t,s+1)} = \text{argmax}_{\beta_j}\mathcal{G}_{\text{PLQ}}(\beta_j; \boldsymbol{\beta}^{(t,s)} \mid \boldsymbol{\beta}^{(t)})$ can be expressed, for instance, as in Theorem 2.2 of Ohishi et al. (2021). For ease of notation, let us consider the one-dimensional function

$$G(r) = -\frac{1}{2}c_2 r^2 + c_1 r - \sum_{j=1}^{m}\sigma_j|r - \delta_j|$$

where $c_2 > 0$ and $\sigma_j > 0$ for all $j = 1, \dots, m$, and define $\{t_a\}_{a=1}^u$ to be the order statistics of $\{\delta_j\}_{j=1}^m$, with $u < m$ if the original set contains duplicates, while $t_0 = -\infty$ and $t_{u+1} = \infty$. Furthermore, let us introduce the auxiliary variables

$$R_a = \begin{cases} (t_a, t_{a+1}] & 0 \leq a \leq u-1 \\ (t_a, t_{a+1}) & a = u \end{cases} \qquad \tilde{\sigma}_a = \begin{cases} -\sum_{j=1}^r \sigma_j & a = 0 \\ \tilde{\sigma}_{a-1} + 2\sum_{j:t_j \in R_a} \sigma_j & 1 \leq a \leq u \end{cases}$$

and $\tilde{r}_a = (c1 - \tilde{\sigma}_a)/c2$ for $a = 1, \dots, u$. Accordingly, it can be shown that it exists either an index a^* such that $\tilde{r}_{a^*} \in R_{a^*}$ or and index $a^\star$ such that $t_{a^\star} \in [\tilde{r}_{a^\star}, \tilde{r}_{a^\star - 1})$ (Ohishi et al., 2021), and that

$$\operatorname{argmax}_r G(r) = \begin{cases} \tilde{r}_{a^*} & \text{if } a^* \text{ exists} \\ t_{a^\star} & \text{otherwise .} \end{cases}$$

In other words, the above solution can be regarded as arising from a generalization of the soft-threshold function of equation (2.16), that notably requires sorting an $n+1$-dimensional array of effective residuals $\{\delta_j\}_{j=1}^m$ appearing within the absolute value functions.

We conclude this Section with a note on the numerical stability of optimization procedures exploiting the PLQ bound, concerning in particular the evaluation of the coefficients in equation (2.18) as their argument goes to zero. Indeed, it appears that state-of-the-art numerical libraries incur into a fictitious oscillating behavior in calculating the function $\big(\log\cosh(r)\big)/r$ in a neighborhood of the origin. We obviate this problem by replacing the aforementioned function with its expansion around zero, whenever its argument is within a ball of arbitrarily small radius around the origin.

Addressing convergence of the corresponding coordinate-wise updates

The approach described above would correspond to an extension of the methodology employed by Wu & Lange (2008) in the context of lasso penalized least-absolute regression, sharing in particular the same limitations. In fact, it has been shown that in such scenarios coordinate-wise optimization may fail to convergence to a proper optimum but rather get stuck into an inferior point (Li & Arce, 2004), since the non-differentiable contribution in the target function is not anymore a separable function of the coordinates of the parameter vector $\boldsymbol{\beta}$ (Tseng, 2001; Hastie et al., 2015), contrarily to the penalty in equation (2.13). Indeed, the same issue is faced by all generalized lasso penalties, which motivated the development of different optimization schemes tailored for the specific penalization matrix $\mathbf{D}$ considered. However, despite the aforementioned connection between the PLQ bound and generalized lasso regularizations, the extension of methodologies developed for the latter to the surrogate in equation (2.21) remains nontrivial. Indeed, optimization schemes dealing with generalized lasso penalties typically leverage on specific sparse structure and discrete nature of the matrix $\mathbf{D}$. Furthermore, such routines often resort to path-wise opti-

mization (Yu et al., 2015; Arnold & Tibshirani, 2016; Ohishi et al., 2021), taking advantage of explicit relations between the solutions of the penalized optimization problems corresponding to different values of a multiplicative penalty parameter λ, which regulates the strength of $\|\mathbf{D}\boldsymbol{\beta}\|_1$ relative to the unconstrained objective as in equation (2.11). On the contrary, while the non-smooth term in equation (2.21) lacks such tunable penalty parameter, the structural constrain among the coordinates of $\boldsymbol{\beta}$ induced by the matrix $\mathbf{D} = \mathbf{N}_{\text{PLQ}}^{(t)}\mathbf{X}$ within the PLQ bound is substantially more complicated than the one appearing in usual generalized lasso regularizations.

Nevertheless, it is worth highlighting that the aforementioned unreliable convergence issue is different in nature from that of Newton's method. Indeed, coordinate-wise optimization schemes targeting the combination $\mathcal{F}(\boldsymbol{\beta})$ of a smooth objective with a non-separable non-differentiable penalty are still guaranteed to reach convergence in a finite number of iterations, intended as observing a difference $\mathcal{F}(\boldsymbol{\beta}^{(t+1)}) - \mathcal{F}(\boldsymbol{\beta}^{(t)})$ smaller than any fixed threshold. On the contrary, Newton's iterates may not reach convergence at all, as they remain prone to face oscillating and even diverging behaviors (Böhning & Lindsay, 1988). Nonetheless, the limiting value arising from coordinate-wise optimization of the PLQ bound may be sub-optimal compared to a proper optimum (Li & Arce, 2004), albeit often barely appreciable in scenarios of practical interest. Finally, we note that coordinate-wise optimization of equation (2.21) reintroduces a trade-off between the enhanced speed of the convergence and the reduced computational efficiency of the resulting MM scheme, if compared to the one arising from the use of quadratic tangent bounds. Indeed, solving the corresponding univariate optimization problems requires a sorting operation with $O(n \log n)$ cost, which is missing when dealing with purely quadratic minorizers. As such, the overall relative performance of the MM schemes exploiting respectively the PLQ and the PG bound will again depend on a balance between cost per update and speed of convergence, the latter being strongly influenced by the dimensionality of the regression problem. Intuitively, the use of the PLQ bound is expected to be advantageous in large-p-small-n settings.

2.4.3 Generalized-MM via semi-smooth coordinate descend

In this Section we propose a hybrid optimization procedure to reduce the computational difficulty of the coordinate-wise updates for the PLQ bound, while still allowing to benefit from the resulting faster rate of convergence. A viable approach to do so would be to get rid of the non-differentiability within the bound, in the search for a compromise between tractability of the surrogate and tightness of the resulting approximation. For instance, this could be done by appealing to the alternative interpretation of the PLQ bound as the combination of least-squares and least-absolute regression terms, which allows for a direct extension of the semi-smooth approximation exploited by Yi & Huang (2017) in the context of quantile regression under elastic net penalty. In particular, the strategy proposed therein amounts to performing an additional approximation step, replacing the absolute value

functions within the PLQ bound with Huber-loss-like contributions, arbitrarily close to the original surrogate

$$|r| \leq d_\epsilon(r) = \begin{cases} |r| & \text{if} \quad |r| > \epsilon \\ \frac{1}{2}\frac{r^2}{\epsilon} + \frac{1}{2}\epsilon & \text{if} \quad |r| \leq \epsilon \end{cases} \qquad \forall \epsilon > 0 \; .$$

Indeed, such operation restores first-order differentiability of the approximate target while still providing a valid MM scheme, as the overall surrogate yields a proper tangent minorizer for the true objective function at $\boldsymbol{\beta}^{(t)}$, provided that $\epsilon < \min_i |\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t)}|$

$$\begin{aligned} \mathcal{G}^{(\epsilon)}_{\text{PLQ}}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) = -\frac{1}{2}\sum_{i=1}^{n} w^{(t)}_{\text{PLQ},i} \cdot \left(\mathbf{x}_i^\intercal \boldsymbol{\beta} - \tau^{(t)}_{\text{PLQ},i}\right)^2 - \sum_{i=1}^{n} \nu^{(t)}_{\text{PLQ},i} \cdot d_\epsilon(\mathbf{x}_i^\intercal \boldsymbol{\beta}) \\ - \lambda(1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}\|_2^2 - \lambda\alpha\|\boldsymbol{\beta}\|_1 + \text{const} \; . \end{aligned} \tag{2.23}$$

Alternatively, it would be possible to obtain a different smoother relaxation of the problem via the relation $|r| < \sqrt{r^2 + \epsilon^2}$ (Voronin et al., 2015), even though strictly speaking the resulting surrogate will not be anymore tangent to the target at $\boldsymbol{\beta}^{(t)}$. However, the resulting iterative optimization scheme could be regarded as a generalization of the MM rationale relaxing the tangency constraints on the minorizer, as the one considered by Parizi et al. (2019). In particular, the authors provided general conditions for the convergence properties of the resulting routines.

In either case, while the smooth approximations would eliminate the hurdle arising from the non-differentiability in the surrogate target, thus restoring once more the computational efficiency of the coordinate-wise updates by getting rid of the sorting operation, the solutions to the associated coordinate-wise optimization problems would not be available anymore via an exact formulation. For this reason, Yi & Huang (2017) suggested solving sequentially the resulting univariate optimizations via semi-smooth Newton coordinate descent, claiming in particular that the convergence of resulting updates is ensured under the pathwise optimization framework. While it may sound contradictory to re-introduce Newton's method to perform the inner optimizations of MM and EM schemes, as one distinctive feature of the latter is indeed the associated convergence guarantees, the empirical studies presented in the following Section suggest that the smooth approximations of the piece-wise linear-quadratic surrogate are sufficiently well behaved for the inner Newton updates to consistently achieve convergence. In particular, this remains true even in specific low-dimensional unpenalized logistic regression settings, where the application of Newton's method to the exact objective actually incurs in the aforementioned diverging behavior. In Appendix A.2.2 we report the coordinate-wise updates for the semi-smooth surrogate from equation (2.23), which follows directly from an extension of the construction by Yi & Huang (2017).

2.5 Empirical studies

In the current Section, we report the results of empirical studies we performed for two antithetical scenarios. We first consider the large-n-small-p setting of the Indian Liver Patient Dataset available at the UCI machine learning repository (Dua & Graff, 2017). Additionally, we consider the large-p-small-n scenario of the `leukemia` Dataset available in the `R` library `supclust`. We perform penalized maximum likelihood estimation for logistic regression under elastic net penalty, where the parameter α in equation (2.11) is set to $\alpha = 0.5$, while the value λ is set by cross-validation via the `glmnet` `R` package on each Dataset. We focus on coordinate-wise optimization via MM (or EM) schemes employing the uniform quadratic bound by Böhning & Lindsay (1988), the optimal quadratic one by Jaakkola & Jordan (2000) and the novel piece-wise linear quadratic bound introduced above. Table 2.1 reports the primary quantity of interest for our work, namely the number of iterations needed to reach convergence, which give empirical evidence on the respective convergence rate. For each lower bound, we consider two possible implementations, which differ in the internal maximization of the corresponding MM schemes. In one case, for each E-step we perform a full inner optimization, until convergence of the coordinate-wise updates (*Full*). Alternatively, for each E-step we perform only one single optimization step per each covariate *(One-step)*, as in generalized EM schemes. For what concern the optimization of the PLQ bound, we proceed by solving exactly the corresponding coordinate-wise updates, as described in Section 2.4.2. On the contrary, we do not explicitly report analogous results for the hybrid strategy discussed in Section 2.4.3, for which the convergence rates are slightly sub-optimal compared to that of the exact coordinate-wise optimization.

Table 2.1: Outcome of penalized logistic regression under elastic net regularization, exploiting different tangent lower bounds of the log-likelihood. In particular, the focus is on the empirical convergence rates of the resulting MM optimization schemes and the associated overall speed-complexity trade-off.

Dataset	λ	n	p	**M-steps**	**Value**	**MM-BL**	**ECM-PG**	**MM-PLQ**
Liver	1,48	579	11	One-step	E steps	750	295	**263**
				Full	E steps	670	177	**131**
					M steps	1592	844	**691**
Leukemia	0,59	72	3572	One-step	E steps	10001	6065	**3156**
				Full	E steps	8827	4836	**2200**
					M steps	10001	5410	**2539**

The results reported in Table 2.1 confirm the intuition reported in the previous Sections. On one hand, the use of the PG bound leads to systematically better performances compared to that arising from the uniform quadratic bound. In fact, the former leads to

consistent faster convergence, and the same cost per iteration. Surprisingly, the PG bound is rarely implemented in state-of-the-art statistical software, such as the aforementioned `glmnet` package, in favor of the uniform one. On the other hand, the use of the PLQ bound in turn improves the convergence rate over the PG one, especially in large-p-small-n settings, where it improves by the same proportions as the latter does over the uniform bound. Despite these encouraging results, the overall speed-complexity trade-off is still in favor of the PG bound, which motivates further investigation for the optimization of the PLQ surrogates, for instance extending the path-wise algorithms developed in the literature for generalized lasso penalties (Tibshirani & Taylor, 2011; Hastie et al., 2015; Arnold & Tibshirani, 2016; Ali & Tibshirani, 2019).

2.6 Discussion

In the present Chapter we reviewed the most popular MM optimization schemes for unpenalized and regularized ML estimation in logistic regression models. We highlighted the optimality of the bound arising from the Pólya-Gamma data augmentation scheme (Polson et al., 2012), among all possible quadratic tangent minorizers for logistic log-likelihood. We stressed how this becomes particularly beneficial in the presence of ℓ_1 penalties, which appears to have been surprisingly overlooked even in state-of-the-art statistical software. Furthermore, we introduced a refined tangent lower bound for logistic log-likelihoods, by providing a piece-wise quadratic tangent minorizer for an intractable term appearing in the latter. Notably, the proposed minorizer dominates over the aforementioned quadratic PG bound, while at the same improves in tractability with respect to alternative piece-wise quadratic minorizers available in the literature (Marlin et al., 2011). As opposed to the latter, the proposed surrogate allows for an explicit analytical expression, regulated by a set of parameters that can be learned adaptively as part of the inferential procedure. The proposed piece-wise quadratic surrogate is particularly suited for the combination with non-smooth penalties, such as the lasso and the elastic net, which already dictate the use of coordinate-wise optimization schemes. In fact, the resulting coordinate-wise updates still admit exact solutions in combination with the PLQ bound. The increased tightness of the novel bound with respect to quadratic surrogates is shown to lead to a faster empirical convergence rate of the resulting MM optimization scheme, especially in high-dimensional scenarios. We note however that the increased convergence speed comes with a trade-off in the cost per iteration, as each coordinate-wise update now requires a sorting operation at $O(n \log n)$ cost, which is absent when dealing with purely quadratic surrogates. Currently, the resulting balance between convergence speed and updates efficiency appears to remain in favor of the PG bound, although the details of the implementation for the PLQ updates might play a pivotal role in assessing such relation. Nevertheless, alternative strategies for the optimization of the proposed PLQ bound certainly deserve further investigation, beyond the aforementioned pathwise optimization schemes, extending those developed

for generalized lasso penalties (Arnold & Tibshirani, 2016). For instance, the impact of the sorting operation can be largely mitigated by considering only a subset of the observations to construct the surrogate target at each step, as it happens in the context of stochastic optimization. Indeed, this would result in a generalization of the MM rationale, in the same spirit as the one studied in Zhang et al. (2019).

Chapter 3

Enhanced variational Bayes for logistic regression via piece-wise quadratic approximations

Variational Bayes routines provide a popular class of strategies to perform approximate posterior computations, whenever a faster alternative to sampling-based Bayesian inference is required. The essence of such methods lies in the minimization of a suitable discrepancy between the exact posterior and an approximate one, belonging to a pre-specified family of tractable distributions. The latter is typically identified by the enforcement of an explicit analytical form or via the imposition of a specific dependence structure in the target parameter space. The choice of the approximating class is often driven by an implicit trade-off between tractability, which ensures the computational advantage over sampling methods, and flexibility, which eventually allows for accurate approximation. In the case of Bayesian logistic regression, several contributions in the literature resort to a Gaussian approximation for the intractable posterior, originally derived by direct tangent minorization of the log-likelihood contributions. Such procedure is actually endowed with a full probabilistic interpretation, as it arises as a proper mean-field (MF) variational Bayes routine under the celebrated Pólya-Gamma data augmentation scheme. As the resulting approximate posterior might not be sufficiently accurate in high dimensions, we leverage on the piecewise linear-quadratic bound derived in the previous Chapter to construct a refined variational approximation of the posterior distribution in logistic regression models, which improves over the PG MF one as a consequence of relative tightness of the corresponding likelihood lower bounds. The novel approximate procedure still allows for simple expression of the update equations for the associated variational parameters, albeit they entail the evaluation of suitable expected values with respect to a distribution with piece-wise quadratic log-kernel, lacking closed-form expressions as opposed to the purely Gaussian case. Nonetheless, it is possible to leverage a well-known scale mixture of normals representation for the Laplace contributions appearing in the lower bound, exploited in the

literature dealing with the Bayesian lasso and quantile regression. This enables to obtain accurate Monte Carlo estimates of the desired quantity, while allowing for an implementation with linear cost in the number of covariates. As a consequence, the resulting approximation still leads to a beneficial tractability-accuracy trade-off in large-p-small-n scenarios, where state-of-the-art exact sampling schemes often face severe limitations while mean-field variational Bayes might suffer from reduced accuracy.

3.1 Introduction

A recurrent hurdle in Bayesian statistics dwells in the need to perform computations under intractable posterior distributions, lacking closed-form expressions. Indeed, the intractability of the posterior often arises directly from a non-conjugate combination of likelihood and prior, or indirectly as a consequence of the dimensionality of the associated statistical problem. The more traditional way to overcome such limitations appeals to different Monte Carlo (MC) schemes tailored to sample from the exact posterior distribution, ranging from most simple methods, such as Accept-Rejection and Importance Sampling (Chopin & Papaspiliopoulos, 2020), to more elaborate Markov Chain Monte Carlo MCMC schemes, including among others Gibbs Samplers, Adaptive Metropolis (Haario et al., 2001), Hamiltonian Monte Carlo (Hoffman & Gelman, 2014). Nonetheless, even most advanced sampling schemes often remain computationally restrictive, if not prohibitive, especially in high-dimensional scenarios or dealing with large datasets and even for moderately complex models (Chopin & Ridgway, 2017; Johndrow et al., 2018). As an alternative way to perform scalable posterior computations, fast deterministic approximations have become increasingly popular in the last two decades, with variational Inference playing a prominent role among several alternatives (Bishop, 2006; Blei et al., 2017). Generally speaking, the term variational Inference encompasses any procedure that replaces the intractable posterior with an approximate one, chosen by optimizing a suitable divergence between the exact target and the elements of a pre-specified family of distributions (Wainwright & Jordan, 2008). Most common variational Bayes (VB) routines focus on the minimization of the forward Kullback-Leibler divergence (Kullback & Leibler, 1951) $\mathrm{KL}[q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} \mid \mathbf{y})]$ over a given family of distribution $q(\boldsymbol{\theta}) \in \mathcal{Q}$, where $\boldsymbol{\theta}$ is the parameter of interest and $\mathbf{y}$ represent the observed data (Ormerod & Wand, 2010; Blei et al., 2017). Nonetheless, the general definition above encompasses a broad spectrum of methodologies focused on the optimization of alternative objectives, ranging from expectation–propagation (Minka, 2001) and belief propagation (Yedidia et al., 2000), to variational routines building on other kinds of divergences, such as the Hellinger distance (Campbell & Li, 2019) or other $\alpha-$Renyi divergences (Li & Turner, 2016), or on direct tangent minorization of the likelihood contributions (Jaakkola & Jordan, 2000; Wainwright et al., 2005).

The different forms of variational inference typically face a common accuracy-scalability trade-off, as considering larger and flexible approximating families will hopefully allow to

get closer to the true objective, most often at the expense of a more elaborate and costly optimization procedure (Bishop, 2006). On the contrary, enforcing simpler structures might decrease the approximation accuracy in exchange for facilitating the optimization process and improving scalability. For instance, mean-field variational inference has become particularly popular thanks to its clear probabilistic interpretation and remarkable scalability, as the underlying independence assumption in the target parameter space allows for efficient coordinate-wise optimization in the case of conditionally conjugate models within the exponential family (Blei et al., 2017). This approach has been particularly exploited in hierarchical formulations with global-local variables, where the unknown parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{z})$ can be split into a set of global target parameters $\boldsymbol{\beta} \in \Re^p$, which are shared across all observations, and a set of local hidden variables $\mathbf{z} = (z_1, \dots, z_n)^\intercal$, each one specific of the single statistical unit (Ormerod & Wand, 2010). Nonetheless, the factorization assumption in MF-VB is known to often lead to an over-simplified approximation of the full target, suffering often from substantial variance underestimation (Giordano et al., 2015), occasionally coupled with over-shrinking of the posterior mean (Fasano et al., 2022). This led to several lines of research improving variational Bayes routines in different directions, often intertwined with one another. Some contributions focused on developing general applicable variational approximation schemes (Ranganath et al., 2014), as opposed to highly model-specific procedures that require non-trivial domain-knowledge and dedicated implementations. Other works combined the essence of mean-field approximations with powerful ideas from gradient-based stochastic optimization (Hoffman et al., 2013; Hoffman & Blei, 2015) and automatic differentiation (Kucukelbir et al., 2017), with the goal of extending scalability via subsampling. Concurrently, other contributions focused on relaxing the independence assumption of MF procedures, in an effort to improve the optimization accuracy both within general formulations (Guo et al., 2016; Miller et al., 2017; Campbell & Li, 2019) or focusing on specific but fundamental statistical constructions (Fasano et al., 2022).

The current Chapter falls within the latter research line, as we develop a refined variational approximation for the posterior distribution under logistic likelihoods, improving over the mean-field VB routine associated with the celebrated PG data augmentation scheme (Polson et al., 2012; Durante & Rigon, 2019). Following the original construction by Jaakkola & Jordan (2000), such MF-VB procedure has been largely exploited in the literature (Bishop & Svensén, 2003; Rasmussen & Williams, 2006; Lee et al., 2010; Ren et al., 2011; Carbonetto & Stephens, 2012), although the formal probabilistic connection with the DA was established only recently thanks to the contribution by Durante & Rigon (2019). We improve over the aforementioned approximation by tackling again the problem from a tangent minorization perspective (Wu & Lange, 2010), exploiting the novel piece-wise quadratic (PLQ) bound derived in the previous Chapter. Indeed, such minorization induces a sharper lower bound on the exact marginal likelihood, dominating over the evidence lower bound (ELBO) associated with the PG MF-VB scheme.

Similarly to the latter, the PLQ variational surrogate involves a set of variational param-

eters $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)$, which can be learned adaptively to optimize the corresponding ELBO. Notably, this still leads to simple expressions for the sequential updates of the aforementioned parameters, although the increased approximation accuracy comes at the cost of reduced tractability, since the purely Gaussian behavior resulting from the MF-VB assumption no longer holds true. However, the specific analytical structure of the log-likelihood minorizers allows to perform computations by borrowing well-established results from the literature, developed in the context of the Bayesian version of lasso (Park & Casella, 2008; Hans, 2009) and quantile regressions (Kozumi & Kobayashi, 2011; Li et al., 2010). In particular, we take advantage of a scale-mixture of Gaussian representation for Laplace random variables, to deal with the piece-wise linear contributions appearing in the approximate log-posterior. This translates into the possibility to produce MC estimates of the desired quantity, sampling from n additional latent variables $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_n)$. Notably, the computational cost given any sample $\boldsymbol{\kappa}$ can be optimized so as to scale only linearly in the number of covariates. This is most interesting in light of the fact that PG MF-VB suffers mostly from reduced accuracy in scenarios with non-negligible posterior skewness, as indeed happens routinely in large-p-small-n settings. As a result, the novel approximation procedure still allows for a beneficial accuracy-tractability trade-off in high dimensions, where a middle-ground between exact but expensive sampling schemes and fast but poor approximations is still lacking in the literature. Preliminary empirical results confirm the potential benefits of the proposed procedure over the PG MF-VB, while more stringent analytical results may follow by extending some recent results in the literature, tackling the PG MF-VB from a tangent transform perspective (Ghosh et al., 2022).

3.2 Mean-field variational Bayes for logistic regression

Bayesian logistic regression notoriously allows for a convenient hierarchical formulation via Pólya-Gamma data augmentation (Polson et al., 2012), as discussed previously in Section 2.2.2

$$
\begin{aligned}
(y_i \mid x_i, \boldsymbol{\beta}) &\overset{\text{ind}}{\sim} \text{Bern}(\pi(\mathbf{x}_i^\intercal \boldsymbol{\beta})) && i = 1, \dots, n \\
(z_i \mid x_i, \boldsymbol{\beta}) &\overset{\text{ind}}{\sim} \text{PG}(1, \mathbf{x}_i^\intercal \boldsymbol{\beta}) && i = 1, \dots, n \\
\boldsymbol{\beta} &\sim \text{N}_p(\boldsymbol{\xi}_0, \boldsymbol{\Omega}_0) \ ,
\end{aligned}
$$

where as before $\pi(\mathbf{x}_i^\intercal \boldsymbol{\beta}) = (1 + e^{-\mathbf{x}_i^\intercal \boldsymbol{\beta}})^{-1}$ represents the logit link, with $i \in \{1, \dots, n\}$, while from now on we omit the predictors $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\intercal$ in the conditioning for ease of notation. A key advantage of such representation lies in the restored conjugacy of the full conditional posterior $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{z})$ under the commonly assumed Gaussian prior for the unknown coefficients $\boldsymbol{\beta}$. While this renewed tractability has been readily exploited within a plain Gibbs Sampling scheme for drawing samples from the exact posterior (Polson et al., 2012), the same representation leads to an efficient mean-field variational Bayes scheme

for approximating the intractable joint posterior $p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})$, as highlighted recently in Durante & Rigon (2019). Such a procedure aims at minimizing the KL divergence between a surrogate and the exact posterior

$$\min_{q(\boldsymbol{\beta},\mathbf{z}) \in \mathcal{Q}_{\text{MF}}} \text{KL}\left[\, q(\boldsymbol{\beta}, \mathbf{z}) \,\|\, p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y}) \,\right],$$

while constraining the former to belong to a factorized family of distributions $\mathcal{Q}_{\text{MF}} = \{\, q(\boldsymbol{\beta}, \mathbf{z}) \mid q(\boldsymbol{\beta}, \mathbf{z}) = q(\boldsymbol{\beta})\, q(\mathbf{z}) \,\}$. More conveniently, the same optimization problem can be recast as the equivalent maximization of the evidence lower bound, allowing to cancel out the unknown and intractable marginal likelihood $p(\mathbf{y}) = \int p(\boldsymbol{\beta}) \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\beta})\, d\boldsymbol{\beta}$

$$\begin{aligned} \text{ELBO}[q(\boldsymbol{\beta}, \mathbf{z})] &= \int q(\boldsymbol{\beta}, \mathbf{z}) \log \frac{p(\boldsymbol{\beta}, \mathbf{z}, \mathbf{y})}{q(\boldsymbol{\beta}, \mathbf{z})} \\ &= -\text{KL}\left[\, q(\boldsymbol{\beta}, \mathbf{z}) \,\|\, p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y}) \,\right] + \log p(\mathbf{y}) \leq \log p(\mathbf{y}). \end{aligned} \tag{3.1}$$

Under the mean-field assumption, such optimization is typically solved via an efficient coordinate-wise optimization scheme (CAVI) (Blei et al., 2017), which translates into the simple sequential updates

$$\begin{aligned} q^{(t+1)}(\boldsymbol{\beta}) &\propto \exp\left\{\, \mathbb{E}_{q^{(t)}(\mathbf{z})}[\log p(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})] \,\right\} \\ q^{(t+1)}(z_i) &\propto \exp\left\{\, \mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})}[\log p(z_i \mid \boldsymbol{\beta})] \,\right\} \qquad i = 1, ..., n\,, \end{aligned}$$

where the additional factorization $q(\mathbf{z}) = \prod_{i=1}^{n} q(z_i)$ is an indirect consequence of the conditional independence structure of the latent variables. Moreover, the augmented representation falls within the tractable framework of conditionally conjugate exponential family models, which allows to further simplify the updates equation above. Indeed, it can be easily verified that $q^{(t+1)}(\boldsymbol{\beta})$ is the density of a p-variate Gaussian $\text{N}_p(\boldsymbol{\xi}^{(t+1)}; \boldsymbol{\Omega}^{(t+1)})$, while $q^{(t+1)}(z_i)$ is the density of a Pólya-Gamma random variable $\text{PG}(1, \zeta_i^{(t+1)})$, where the respective parameters are obtained as

$$\begin{aligned} \boldsymbol{\Omega}^{(t+1)} &= \left(\boldsymbol{\Omega}_0^{-1} + \mathbf{X}^\intercal \text{diag}\big(\{w_{\text{PG}}(\zeta_i^{(t)})\}_{i=1}^{n}\big)\mathbf{X}\right)^{-1} \\ \boldsymbol{\xi}^{(t+1)} &= \boldsymbol{\Omega}^{(t+1)} \left(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\xi}_0 + \mathbf{X}^\intercal(\mathbf{y} - 1/2\, \mathbf{1}_n)\right) \\ \big(\zeta_i^{(t+1)}\big)^2 &= \mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})}\left[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2\right] = \mathbf{x}_i^\intercal \boldsymbol{\Omega}^{(t+1)} \mathbf{x}_i + (\mathbf{x}_i^\intercal \boldsymbol{\xi}^{(t+1)})^2 \qquad i = 1, \ldots, n\,, \end{aligned} \tag{3.2}$$

where $w_{\text{PG}}(\zeta_i^{(t)}) = \mathbb{E}_{q^{(t)}(z_i)}[z_i] = \tanh(\zeta_i^{(t)}/2)/(2\zeta_i^{(t)})$, while $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$. The updates above are repeated iteratively until convergence of the corresponding ELBO, the latter being ensured by a useful monotonicity property for the lower bound itself.

3.2.1 Variational inference via tangent minorization

The same variational approximation was originally introduced from a different perspective in the seminal contribution by Jaakkola & Jordan (2000), without appealing to a formal hierarchical probabilistic interpretation. Indeed, the authors tackled the problem by providing a family of tangent minorizers for the log-likelihood contributions

$$\log p(y_i \mid \boldsymbol{\beta}) \geq \log \bar{p}(y_i \mid \boldsymbol{\beta}, \zeta_i) \qquad \forall \boldsymbol{\beta} \in \Re^p$$
$$\log p(y_i \mid \bar{\boldsymbol{\beta}}) = \log \bar{p}(y_i \mid \bar{\boldsymbol{\beta}}, \zeta_i) \qquad \forall \bar{\boldsymbol{\beta}} = \bar{\boldsymbol{\beta}}(\zeta_i)$$

for $i = 1, \ldots, n$, where the parameters $\{\zeta_i\}_{i=1}^n$ were interpreted as simply identifying tangent locations. Regardless of the specific parametric form of the surrogates $\bar{p}(y_i \mid \boldsymbol{\beta}, \zeta_i)$, the above minorization implicitly induces both a lower bound on the marginal likelihood

$$p(\mathbf{y}) \geq \bar{p}(\mathbf{y} \mid \boldsymbol{\zeta}) = \int p(\boldsymbol{\beta}) \prod_{i=1}^{n} \bar{p}(y_i \mid \boldsymbol{\beta}, \zeta_i) \, d\boldsymbol{\beta} \tag{3.3}$$

and an associated full-fledged approximate posterior distribution

$$\bar{p}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}) = \frac{\bar{p}(\boldsymbol{\beta}, \mathbf{y} \mid \boldsymbol{\zeta})}{\bar{p}(\mathbf{y} \mid \boldsymbol{\zeta})} = \frac{p(\boldsymbol{\beta}) \prod_{i=1}^{n} \bar{p}(y_i \mid \boldsymbol{\beta}, \zeta_i)}{\bar{p}(\mathbf{y} \mid \boldsymbol{\zeta})}. \tag{3.4}$$

In particular, the lower bound on the marginal likelihood in equation (3.3) can be employed as a global measure of the approximation accuracy, hence redirecting the attention to its maximization over the variational parameters $\max_{\boldsymbol{\zeta} \in \Re^n} \bar{p}(\mathbf{y} \mid \boldsymbol{\zeta})$. Such optimization can be performed by resorting to a further minorization via Jensen inequality, which gives

$$\log \bar{p}(\mathbf{y} \mid \boldsymbol{\zeta}) = \log \mathbb{E}_{q(\boldsymbol{\beta})} \left[\frac{\bar{p}(\boldsymbol{\beta}, \mathbf{y} \mid \boldsymbol{\zeta})}{q(\boldsymbol{\beta})} \right] \geq \mathbb{E}_{q(\boldsymbol{\beta})} \left[\log \frac{\bar{p}(\boldsymbol{\beta}, \mathbf{y} \mid \boldsymbol{\zeta})}{q(\boldsymbol{\beta})} \right] . \tag{3.5}$$

While the above relation holds true for any distribution $q(\boldsymbol{\beta})$, it becomes of practical interest by setting $q^{(t+1)}(\boldsymbol{\beta}) = \bar{p}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})$, which gives

$$\mathbb{E}_{\bar{p}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})} \left[\log \frac{\bar{p}(\boldsymbol{\beta}, \mathbf{y} \mid \boldsymbol{\zeta})}{\bar{p}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})} \right] = Q(\boldsymbol{\zeta} \mid \boldsymbol{\zeta}^{(t)}) - H(\boldsymbol{\zeta}^{(t)}) \, ,$$

where $H(\boldsymbol{\zeta}^{(t)})$ represents the Shannon entropy of the distribution $\bar{p}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})$, which does not depend on the free parameters $\boldsymbol{\zeta}$, while $Q(\boldsymbol{\zeta} \mid \boldsymbol{\zeta}^{(t)})$ is defined as

$$Q(\boldsymbol{\zeta} \mid \boldsymbol{\zeta}^{(t)}) = \mathbb{E}_{\bar{p}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})} \left[\log \bar{p}(\boldsymbol{\beta}, \mathbf{y} \mid \boldsymbol{\zeta}) \right] . \tag{3.6}$$

As equation (3.5) now holds as an equality when $\boldsymbol{\zeta} = \boldsymbol{\zeta}^{(t)}$, the minorization above naturally translates into a genuine minorize-maximize (MM) routine (Hunter & Lange, 2004; Wu &

Lange, 2010) for optimizing $\bar{p}(\mathbf{y} \mid \boldsymbol{\zeta})$ over $\boldsymbol{\zeta} \in \Re^n$, by updating the variational parameters sequentially via the fixed-point scheme $\boldsymbol{\zeta}^{(t+1)} = \boldsymbol{\mathcal{T}}(\boldsymbol{\zeta}^{(t)}) = \text{argmax}_{\boldsymbol{\zeta} \in \Re^n} Q(\boldsymbol{\zeta} \mid \boldsymbol{\zeta}^{(t)})$, and iteratively using the updated locations to construct refined tangent minorizers. By virtue of the properties of MM schemes, the monotonicity of $\{\bar{p}(\mathbf{y} \mid \boldsymbol{\zeta}^{(t+1)})\}_{t \geq 0}$ is ensured as well under the above formulation, which in turn guarantees convergence of the procedure under usual assumptions for MM schemes (Wu & Lange, 2010), eventually reaching an equilibrium point $\boldsymbol{\zeta}^* = \boldsymbol{\mathcal{T}}(\boldsymbol{\zeta}^*)$. Furthermore, the joint optimization in $\Re^n$ naturally decouples as n separate one-dimensional maximization problem, as a consequence of the additive nature of the log-likelihood

$$\zeta_i^{(t+1)} = \underset{\zeta_i \in \Re}{\text{argmax}} \; \mathbb{E}_{\bar{p}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})} \left[\log \bar{p}(y_i \mid \boldsymbol{\beta}, \zeta_i) \right] \; . \tag{3.7}$$

It is worth mentioning that the tangent minorization perspective to construct variational approximations of the posterior does not offer guarantees on the KL divergence between the latter and the optimal surrogate, as in general

$$\text{argmax}_{\boldsymbol{\zeta} \in \Re^n} \bar{p}(\mathbf{y} \mid \boldsymbol{\zeta}) \neq \text{argmin}_{\boldsymbol{\zeta} \in \Re^n} \text{KL} \left[\bar{p}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}) \,\|\, p(\boldsymbol{\beta} \mid \mathbf{y}) \right], \tag{3.8}$$

as exemplified later on in Section 3.4.

3.2.2 Quadratic surrogates and equivalence with PG MF-VB

A key step in constructing a practical variational approximation in the setting of equation (3.3) consists in supplying a valid tangent lower bound for a specific likelihood. While elaborate and flexible parametric forms of $\bar{p}(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\zeta})$ might lead to more accurate approximations, simpler surrogates might allow for a more efficient solution to equation (3.7). In the case of logistic regression, Jaakkola & Jordan (2000) focused on quadratic minorizer for the log-likelihood

$$\begin{aligned} \log p(y_i \mid \boldsymbol{\beta}) &= (y_i - 0.5)\mathbf{x}_i^\intercal \boldsymbol{\beta} - \log(e^{\mathbf{x}_i^\intercal \boldsymbol{\beta}/2} + e^{-\mathbf{x}_i^\intercal \boldsymbol{\beta}/2}) \\ &\geq (y_i - 0.5)\mathbf{x}_i^\intercal \boldsymbol{\beta} - \log(e^{\zeta_i/2} + e^{-\zeta_i/2}) - \frac{1}{2} w_{\text{PG}}(\zeta_i)((\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 - \zeta_i^2) \\ &:= \log \bar{p}_{\text{PG}}(y_i \mid \boldsymbol{\beta}, \zeta_i), \end{aligned}$$

obtained via supporting hyperplane inequality, with equality clearly holding when $(\zeta_i)^2 = (\mathbf{x}_i^\intercal \boldsymbol{\beta})^2$. This induces a tractable Gaussian Variational approximation of the posterior $\bar{p}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}) = \phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}(\boldsymbol{\zeta}); \boldsymbol{\Omega}(\boldsymbol{\zeta}))$, which has been indeed largely exploited in the literature (Bishop & Svensén, 2003; Rasmussen & Williams, 2006; Lee et al., 2010; Ren et al., 2011; Carbonetto & Stephens, 2012). As a matter of fact, both the approximate posterior parameters $\boldsymbol{\xi}(\boldsymbol{\zeta})$ and $\boldsymbol{\Omega}(\boldsymbol{\zeta})$ and the update rules of the variational parameters $\boldsymbol{\zeta}^{(t+1)}$ match exactly with that of equation (3.2). Indeed the original tangent minorization approach by Jaakkola & Jordan

(2000) fully coincides with the proper MF-VB under the Pólya-Gamma DA presented in the previous Section, as recently shown by Durante & Rigon (2019). In particular, the lower bound in equation (3.5) coincides with the genuine evidence lower bound for the joint approximate posterior from equation (3.1)

$$\text{ELBO}[q^{(t+1)}(\boldsymbol{\beta}, \mathbf{z})] = \int q^{(t+1)}(\boldsymbol{\beta}, \mathbf{z}) \log \frac{p(\boldsymbol{\beta}, \mathbf{z}, \mathbf{y})}{q^{(t+1)}(\boldsymbol{\beta}, \mathbf{z})} = \mathbb{E}_{\bar{p}_{\text{PG}}(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\zeta}^{(t)})} \left[\log \frac{\bar{p}_{\text{PG}}(\boldsymbol{\beta}, \mathbf{y} \mid \boldsymbol{\zeta}^{(t+1)})}{\bar{p}_{\text{PG}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})} \right],$$

as proven in Theorem 2.1 of Durante & Rigon (2019), implying that the two procedures ultimately aim at maximizing the same objective.

As outlined in the previous Section, this result places the approximation by Jaakkola & Jordan (2000) within the broader framework of variational inference for conditionally conjugate exponential family models, thereby inheriting recent advances derived for this latter class. For instance, Wang & Titterington (2004) proved that the expected value for the parameter $\boldsymbol{\beta}$ under the MF-VB approximation converges locally to the true value with probability 1 as the sample size becomes indefinitely large. In fact, analogous guarantees on the approximation accuracy in large n regimes from a tangent minorization perspective were recently studied in Ghosh et al. (2022), by upper bounding the frequentist risk bound. On the contrary, it is well known that MF-VB procedures as the one above might suffer from variance underestimation (Giordano et al., 2015), and possibly over-shrinking of the posterior mean (Fasano et al., 2022), especially in situations where the exact posterior for $\boldsymbol{\beta}$ exhibits substantial skewness. While in Section 3.4 we give an example of this issue in a low-dimensional setting, the logit posterior $p(\boldsymbol{\beta} \mid \mathbf{y})$ is known to be often significantly skewed in large-p-small-n scenarios, as it happens indeed also under the probit link (Durante, 2019). As for the latter case, the reduced accuracy of the variational approximation typically has a dramatic impact also on prediction performances in such high-dimensional settings (Fasano et al., 2022), motivating the development of refined approximations. In the remaining of the present Chapter, we show how this can be done, for instance, by providing sharper tangent minorizers for the likelihood contributions, thereby covering a gap in the literature.

3.3 Improved variational inference via piece-wise quadratic tangent bounds

The different perspectives on variational Bayes presented above suggest two alternative strategies to construct refined approximations of the exact posterior. An elegant probabilistic approach would be to relax the independence assumption of the mean-field family, as done for instance by Fasano et al. (2022) in the context of probit regression, where the authors considered approximate joint posteriors of the form $q(\boldsymbol{\beta}, \mathbf{z}) = q(\boldsymbol{\beta} \mid \mathbf{z}) \prod_{i=1}^{n} q(z_i)$. However, an analogous relaxation for logistic regression would not be equally tractable

under the PG DA. A viable alternative would be to provide tighter lower bounds for the likelihood terms, exploited with the tangent minorization perspective to variational Bayes presented above. Indeed, this can be done by leveraging on the piece-wise linear quadratic bounds introduced in the previous Chapter, which fits within a broader endeavor in the literature to improve over the quadratic bound for logistic likelihoods by Jaakkola & Jordan (2000) via piece-wise quadratic functions (Khan et al., 2010; Marlin et al., 2011; Ermis & Bouchard, 2014). In particular, Marlin et al. (2011) consider a general but implicit formulation accounting for all continuous piece-wise quadratic functions over an arbitrary partition of the real line, as detailed previously in Section 2.4. For a pre-specified cardinality $R \geq 2$ of such partition, the lower bound was constructed by solving a data-agnostic constrained optimization problem both on the locations identifying the interval's separation and on the local coefficients of the quadratic contributions. The output of such numeric optimization was then exploited within a generalized EM algorithm to overcome the intractability of some logistic-Gaussian integrals, replacing the logistic log-likelihood with a pre-calculated minimax R-piece-wise quadratic bound.

Albeit formally included in such a general formulation, the methodology we propose differs in several aspects from that of Marlin et al. (2011). Above all, we provide an explicit analytical formulation of the tangent bound, while restricting $R = 2$ and fixing the splitting point of the two-fold partition at the origin. At the same time, we allow for heterogeneity in the coefficients for the bound of each likelihood contribution, each regulated via a single scalar parameter ζ_i whose value is learned adaptively from the data. Finally, since the piece-wise behavior originates solely by terms proportional to the absolute values of the linear predictors $|\mathbf{x}_i^\intercal \boldsymbol{\beta}|$, we can take advantage of a well-known scale-mixture of normals representation for Laplace random variables to overcome the reduced tractability compared to purely Gaussian approximations.

3.3.1 Variational Bayes via PLQ tangent minorization

The VB procedure we propose leverage on the novel piece-wise quadratic bound for logistic log-likelihoods, introduced in the previous Chapter

$$\begin{aligned}\log \bar{p}_{\text{PLQ}}(y_i \mid \boldsymbol{\beta}, \zeta_i) &:= (y_i - 0.5)\mathbf{x}_i^\intercal \boldsymbol{\beta} - \log(e^{\zeta_i/2} + e^{-\zeta_i/2}) - \frac{1}{2} w_{\text{PLQ}}(\zeta_i)\left((\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 - \zeta_i^2\right) \\ &\quad - \nu_{\text{PLQ}}(\zeta_i)\left(|\mathbf{x}_i^\intercal \boldsymbol{\beta}| - |\zeta_i|\right) \\ &= (y_i - 0.5)\mathbf{x}_i^\intercal \boldsymbol{\beta} - \frac{1}{2} w_{\text{PLQ}}(\zeta_i)\,(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 - \nu_{\text{PLQ}}(\zeta_i)\,|\mathbf{x}_i^\intercal \boldsymbol{\beta}| - \log 2,\end{aligned} \tag{3.9}$$

where as before

$$w_{\text{PLQ}}(\zeta_i) = 2\, w_{\text{PG}}(\zeta_i) - 2 \log\cosh\left(\zeta_i/2\right)/\zeta_i^2 \qquad \nu_{\text{PLQ}}(\zeta_i) = |\zeta_i|\big(w_{\text{PG}}(\zeta_i) - w_{\text{PLQ}}(\zeta_i)\big).$$

In particular, we have already shown in Section 2.4.1 that $\bar{p}_{\text{PG}}(y_i \mid \beta, \zeta_i)$ acts in turn as a tangent lower bound also for the surrogate in equation (3.9), which implies that

$$\log p(y_i \mid \boldsymbol{\beta}) \geq \log \bar{p}_{\text{PLQ}}(y_i \mid \boldsymbol{\beta}, \zeta_i) \geq \log \bar{p}_{\text{PG}}(y_i \mid \boldsymbol{\beta}, \zeta_i) \qquad \forall \beta \in \Re^p, \tag{3.10}$$

with equalities holding whenever $|\zeta_i| = |\mathbf{x}_i^\intercal \boldsymbol{\beta}|$, for all $i = 1, \dots, n$. Accordingly, by simply plugging-in the enhanced tangent surrogate in equation (3.4), we obtain a more flexible variational approximation for the posterior, with piece-wise quadratic kernel

$$\bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}) \propto \exp\Big\{ -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\xi}_0)^\intercal \boldsymbol{\Omega}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi}_0) + \sum_{i=1}^n (y_i - 0.5)\mathbf{x}_i^\intercal \boldsymbol{\beta} - \frac{1}{2}\sum_{i=1}^n w_{\text{PLQ}}(\zeta_i)\,(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 - \sum_{i=1}^n \nu_{\text{PLQ}}(\zeta_i)\,|\mathbf{x}_i^\intercal \boldsymbol{\beta}| \Big\} .$$

In particular, the relative tightness of the likelihood tangent minorizers transfers directly to the corresponding evidence lower bounds from equation (3.3).

Lemma 3.1. *For any $\zeta \in \Re^n$ it holds that*

$$p(\mathbf{y}) \geq \bar{p}_{\text{PLQ}}(\mathbf{y} \mid \boldsymbol{\zeta}) \geq \bar{p}_{\text{PG}}(\mathbf{y} \mid \boldsymbol{\zeta}).$$

Therefore, defining $\boldsymbol{\zeta}_{\text{PLQ}} = \text{argmax}_{\boldsymbol{\zeta}}\, \bar{p}_{\text{PLQ}}(\mathbf{y} \mid \boldsymbol{\zeta})$ and $\boldsymbol{\zeta}_{\text{PG}} = \text{argmax}_{\boldsymbol{\zeta}}\, \bar{p}_{\text{PG}}(\mathbf{y} \mid \boldsymbol{\zeta})$ one has

$$p(\mathbf{y}) \geq \bar{p}_{\text{PLQ}}(\mathbf{y} \mid \boldsymbol{\zeta}_{\text{PLQ}}) \geq \bar{p}_{\text{PLQ}}(\mathbf{y} \mid \boldsymbol{\zeta}_{\text{PG}}) \geq \bar{p}_{\text{PG}}(\mathbf{y} \mid \boldsymbol{\zeta}_{\text{PG}}).$$

In this sense, the variational approximation via piece-wise quadratic approximation is guaranteed to improve in accuracy over the PG MF-VB one. As previously mentioned, it is not straightforward to translate the results above into quantitative guarantees on the KL divergence or on the accuracy of approximate posterior moments. Nevertheless, the empirical results reported in Section 3.4 show that the refined variational approximation can lead to significant improvements even in these directions.

Update equations for the variational parameters

As before, the optimization of the marginal likelihood lower bound over the parameters $\boldsymbol{\zeta}$ can be performed via an EM scheme, by sequentially maximizing $Q_{\text{PLQ}}(\boldsymbol{\zeta} \mid \boldsymbol{\zeta}^{(t)}) = \mathbb{E}_{\bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})}\left[\log \bar{p}_{\text{PLQ}}(\boldsymbol{\beta}, \mathbf{y} \mid \boldsymbol{\zeta})\right]$. Notably, this still leads to simple update rules for the variational parameters $\boldsymbol{\zeta}^{(t+1)} = \boldsymbol{\mathcal{T}}_{\text{PLQ}}(\boldsymbol{\zeta}^{(t)})$, which boils down to

$$\left|\zeta_i^{(t+1)}\right| = \frac{\mathbb{E}_{\bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})}\left[\,(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2\,\right]}{\mathbb{E}_{\bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})}\left[\,|\mathbf{x}_i^\intercal \boldsymbol{\beta}|\,\right]} \qquad i = 1, \dots, n \tag{3.11}$$

as proved in Appendix A.3.1. As the purely Gaussian behavior is lost under the PLQ approximation, the quantities above lack of closed-form expressions. However, the proposed variational approximation can still be beneficial in high-dimensional settings, as the specific nature of the involved log-likelihood minorizers allows to exploit a well-known scale-mixture representation for the Laplace-like contributions to obtain accurate MC estimates of the required expected values. At the same time, we propose a generalized-MM scheme (Parizi et al., 2019) for the optimization of $\bar{p}_{\text{PLQ}}(\mathbf{y} \mid \boldsymbol{\zeta})$ over $\boldsymbol{\zeta} \in \Re^n$ that results in a hybrid approximation, benefiting both from the tractability of the PG MF-VB one and the more accurate objective induced by the PLQ bound.

3.3.2 Efficient MC estimates via scale-mixture representation

Laplace random variables admit a well-known scale-mixture of normals representation, arising from the integral equation

$$\frac{a}{2}\, e^{-a|r|} = \int_0^\infty \frac{1}{\sqrt{2\pi\kappa}}\, e^{-r^2/2\kappa}\, \frac{a^2}{2}\, e^{-a^2\kappa/2} d\kappa \,, \tag{3.12}$$

which has been previously exploited in the literature within Gibbs Sampling schemes for drawing from the posterior of the Bayesian version of lasso (Park & Casella, 2008; Hans, 2009) and quantile regressions (Kozumi & Kobayashi, 2011; Li et al., 2010). Since the piecewise behavior in the PLQ minorizers originates solely by terms proportional to the absolute values of the linear predictors $|\mathbf{x}_i^\intercal \boldsymbol{\beta}|$, we can take advantage of such property to overcome the reduced tractability of the corresponding variational approximation. Indeed, by conditioning on the latent variables $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_n)^\intercal$ associated with the representation in equation (3.12), the PLQ bound becomes again a purely quadratic function of $\boldsymbol{\beta}$, allowing for closed-form computations conditionally on $\boldsymbol{\kappa}$. This naturally suggests a way to calculate MC estimates of the desired quantities $\mathbb{E}\left[f(\boldsymbol{\beta})\right] = \mathbb{E}\left[\mathbb{E}\left[f(\boldsymbol{\beta}) \mid \boldsymbol{\kappa}\right]\right]$. For instance, the marginal likelihood lower bound can be rewritten as

$$\bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}) = \frac{1}{2^n} \int_{(\Re^+)^n} \frac{\phi_p\big(\boldsymbol{\xi}_0; \boldsymbol{\Omega}_0\big)}{\phi_p\big(\boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa}); \boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa})\big)} \prod_{i=1}^n p(\kappa_i)\, d\kappa_i = \frac{1}{2^n}\, \mathbb{E}_{p(\boldsymbol{\kappa})}\left[\, \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \,\right] \tag{3.13}$$

where the $\boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa})$, $\boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa})$ and $\varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa})$ are defined as

$$\begin{aligned}
\boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa}) &= \Big(\boldsymbol{\Omega}_0^{-1} + \mathbf{X}^\intercal \text{diag}\big(\{w_{\text{PLQ}}(\zeta_i) + \nu^2_{\text{PLQ}}(\zeta_i)/\kappa_i\}_{i=1}^n\big)\mathbf{X}\Big)^{-1} \\
\boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa}) &= \boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa})\, \Big(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\xi}_0 + \mathbf{X}^\intercal\big(\mathbf{y} - 0.5\, \mathbf{1}_n\big)\Big) \\
\varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa}) &= \phi_p\big(\boldsymbol{\xi}_0; \boldsymbol{\Omega}_0\big) / \phi_p\big(\boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa}); \boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa})\big) \,,
\end{aligned}$$

while $p(\kappa_i) = e^{-\kappa_i/2}/\sqrt{2\pi\kappa_i}$ are the densities of Chi-squared random variables with one degree of freedom. Similarly, the updates for the variational parameters in equation (3.11) take the form

$$|\zeta_i^{(t+1)}| = \frac{\mathbb{E}_{p(\boldsymbol{\kappa})}\left[\, \varpi(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa}) \cdot \left(\sigma_i^2(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa}) + \mu_i^2(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa})\right) \right]}{\mathbb{E}_{p(\boldsymbol{\kappa})}\left[\frac{\varpi(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa})}{\sigma_i(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa})} \cdot \left(2\,\phi\Big(\frac{\mu_i(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa})}{\sigma_i(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa})}\Big) + \frac{\mu_i(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa})}{\sigma_i(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa})}\Big(1 - 2\,\Phi\Big(-\frac{\mu_i(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa})}{\sigma_i(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa})}\Big)\Big)\right)\right]}$$

where $\mu_i(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa}) = \mathbf{x}_i^\intercal \boldsymbol{\xi}(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa})$ and $\sigma_i^2(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa}) = \mathbf{x}_i^\intercal \boldsymbol{\Omega}(\boldsymbol{\zeta}^{(t)}, \boldsymbol{\kappa})\mathbf{x}_i$.

We highlight that the above sampling strategy would still be more convenient than standard MCMC methods in high dimensions, as it entails $i.i.d.$ sampling and leads to efficient computations, since the bottleneck step can be made to scale linearly in p. Accordingly, equation (3.12) would allow to exploit successfully the proposed methodology in large-p-small-n scenarios, where exact sampling schemes are inefficient while the PG MF-VB can be inaccurate. We refer to Appendix A.3.2 for analogous expressions for posterior moments and for the details on the scalable calculations of the weights $\varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa})$.

3.3.3 Hybrid PG VB via generalized EM optimization

Notwithstanding the possibility to exploit the aforementioned scale-mixture representation to deal with the Laplace contributions, computations under the PLQ variational approximation remain essentially more involved than those entailed by plain quadratic surrogates, as in the PG MF-VB approach. For this reason, we hereby propose a hybrid strategy that benefits from the best of both approaches, as it preserves the tractability of the Gaussian approximation arising from the bound by Jaakkola & Jordan (2000), while still taking advantage from the refined objective function induced by the PLQ bound. The approach we propose can be formally regarded as a generalization of traditional EM schemes, where the E-step is approximated by performing the required calculation under an alternative distribution. The underlying idea traces back to the so-called incremental-EM by Neal & Hinton (1998) and the variational-EM by Jordan et al. (1999), later extensively addressed by Gunawardana & Byrne (2005) in the framework of generalized alternating minimization (GAM). More recently, Parizi et al. (2019) extended the same rationale to the broader setting of MM schemes, relaxing the tangency condition for the lower bounds via the so-called generalized majorization-minimization (G-MM).

Generalized alternating minimization framework

Indeed, the minorization in equation (3.5) holds true regardless of the specific form of the proxy distribution $q(\boldsymbol{\beta})$, while the specific choice $q^{(t+1)}(\boldsymbol{\beta}) = \bar{p}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})$ is essentially driven by the need to construct a lower bound that is tangent to the objective at the current locations. The latter condition is in fact a fundamental requirement of traditional MM and

EM optimization schemes, as it is crucial in ensuring the monotonicity of the corresponding updates. However, one can proceed by relaxing such tangency condition and performing the expectation in equation (3.6) under a different, more tractable distribution, while keeping the argument of the expected value fixed to $\log \bar{p}_{\text{PLQ}}(\boldsymbol{\beta}, \mathbf{y} \mid \boldsymbol{\zeta})$. In particular, we propose to use as surrogate distribution $q^{(t+1)}(\boldsymbol{\beta}) = \bar{p}_{\text{PG}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t+1/2)})$, for some suitable $\boldsymbol{\zeta}^{(t+1/2)}$. The same hybrid strategy arises by addressing EM schemes from an alternating minimization perspective, according to which also the E-step is described as solving an optimization problem

$$q^{(t+1)}(\boldsymbol{\beta}) = \underset{q(\boldsymbol{\beta})}{\operatorname{argmin}}\, \text{KL}\big[\, q(\boldsymbol{\beta}) \,\|\, \bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)}) \,\big] = \bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})\,.$$

Accordingly, in generalized alternating minimization schemes the above optimization is relaxed by constraining $q^{(t+1)}(\boldsymbol{\beta})$ to belong to a given subclass of distributions

$$q^{(t+1)}(\boldsymbol{\beta}) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}}\, \text{KL}\big[\, q(\boldsymbol{\beta}) \,\|\, \bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)}) \,\big]\,,$$

where in particular we choose the parametric family arising from the PG MF scheme $\mathcal{Q} = \big\{\, \bar{p}_{\text{PG}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}) \mid \boldsymbol{\zeta} \in \Re^n \,\big\}$. This leads to a set of hybrid updates for the variational parameters

$$\begin{aligned}
\big|\zeta_i^{(t+1)}\big| = \big[\boldsymbol{\mathcal{T}}_{h\text{PG}}(\boldsymbol{\zeta}^{(t)})\big]_i &= \frac{\mathbb{E}_{\bar{p}_{\text{PG}}(\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\zeta}^{(t+1/2)})}\big[\,(\mathbf{x}_i^\intercal\boldsymbol{\beta})^2\,\big]}{\mathbb{E}_{\bar{p}_{\text{PG}}(\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\zeta}^{(t+1/2)})}\big[\,|\mathbf{x}_i^\intercal\boldsymbol{\beta}|\,\big]} \\
&= \frac{(\tilde{\sigma}_i^{(t+1)})^2 + (\tilde{\mu}_i^{(t+1)})^2}{2\,\tilde{\sigma}_i^{(t+1)}\,\phi\big(\tilde{\mu}_i^{(t+1)}/\tilde{\sigma}_i^{(t+1)}\big) + \tilde{\mu}_i^{(t+1)}\Big(1 - 2\,\Phi\big(-\tilde{\mu}_i^{(t+1)}/\tilde{\sigma}_i^{(t+1)}\big)\Big)}\,,
\end{aligned}$$

where we have defined $\boldsymbol{\zeta}^{(t+1/2)} = \operatorname{argmin}_{\boldsymbol{\zeta}} \text{KL}[\,\bar{p}_{\text{PG}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}) \,\|\, \bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})\,]$, $\tilde{\mu}_i^{(t+1)} = \mathbf{x}_i^\intercal\tilde{\boldsymbol{\xi}}^{(t+1)}$ and $\tilde{\sigma}_i^{(t+1)} = (\mathbf{x}_i^\intercal\tilde{\boldsymbol{\Omega}}^{(t+1)}\mathbf{x}_i)^{1/2}$, while $\tilde{\boldsymbol{\xi}}^{(t+1)}$ and $\tilde{\boldsymbol{\Omega}}^{(t+1)}$ are obtained as in equation (3.2), but replacing $\boldsymbol{\zeta}^{(t)}$ with $\boldsymbol{\zeta}^{(t+1/2)}$. The sequential repetition of the above update rules thereby formally corresponds to a GAM scheme, aiming at the approximate maximization over $\boldsymbol{\zeta}$ of the ELBO $\bar{p}_{\text{PLQ}}(\mathbf{y} \mid \boldsymbol{\zeta})$ via non-tangent minorizer of the latter, induced by the family of quadratic surrogates defined by Jaakkola & Jordan (2000). As a side effect, this ultimately produces also a refined Gaussian approximation of the posterior, which preserves the analytical form arising from the mean-field scheme $\bar{p}_{\text{PG}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta})$, but optimizes the associated variational parameters $\boldsymbol{\zeta}$ as to maximize a refined surrogate objective, induced by the PLQ bound. Indeed, preliminary empirical results suggest that this strategy indirectly puts a remedy for the mean over-shrinking and variance underestimation that affect MF-VB in skewed and high-dimensional scenarios.

Comparison with alternative correction methods for MF-VB

In this respect, it is interesting to notice a practical analogy between the hybrid strategy proposed above and the linear response variational Bayes (LRVB) by Giordano et al. (2015). Indeed the authors introduced a method for correcting the uncertainty underestimation of MF-VB by generalizing linear response methods from statistical physics. The essence of their approach lies in performing a perturbation of the MF fixed-point equation for the variational parameters $\boldsymbol{\zeta} = \boldsymbol{\mathcal{T}}_{\text{MF}}(\boldsymbol{\zeta})$, induced by considering an exponential tilting of the posterior, or equivalently, a log-linear perturbation of the latter. For approximate posterior within the exponential family, this translates into a simple intuitive formula for calculating the linear response correction by solving a linear system based on the MF-VB solution. While elegant and generally applicable, the methodology by Giordano et al. (2015) crucially relies on the assumption that MF-VB provides an accurate approximation of the posterior mean, which in fact might not hold in the high-dimensional settings that we consider in this work. Conversely, the hybrid strategy we propose, which interestingly can be regarded as arsing from a log-piece-wise linear perturbation of the posterior, acts in practice as a correction of the aforementioned location bias. Furthermore, the LRVB approach increases non trivially the computational burden, as it scales linearly with the number of data points but cubically in the number of covariates. On the contrary, our hybrid variational approximation comes with the same exact computational cost as the standard PG MF-VB, while still allowing for the combination with post-processing by Giordano et al. (2015).

3.4 Empirical studies

In this Section, we demonstrate the potential benefit of the proposed PLQ variational approximation for the logistic posterior in a simple simulation study. Albeit low-dimensional, the exact posterior is strongly skewed, leading to poor performances of the PG MF-VB. Indeed, we consider an intercept only logistic regression $y_i \mid \beta \overset{\text{ind}}{\sim} \text{Bern}(\pi(\beta))$ where the true value of the parameter is set to $\beta_{\text{TRUE}} = 3$, while the prior is chosen to be a normal $\text{N}(0, \sigma_0^2)$ with $\sigma_0 = 3.5$. We simulate $n = 10$ observations from the model, which results in $y_i = 1$ for all $i = 1, \dots, n$, where such an extremely unbalanced scenario is indeed purposely designed to obtain a skewed but easily visualized exact posterior. Furthermore, the PLQ variational approximation allows for exact computations in such a low-dimensional setting, leading to the results reported in Figure 3.1 and Table 3.1. It is interesting to notice that, as already mentioned previously, the tangent transform approach offers guarantees solely on the relative tightness of the evidence lower bounds, but not for the corresponding KL divergence from the exact posterior. Indeed, defining as before $\zeta_{\text{PLQ}} = \text{argmax}_\zeta \, \bar{p}_{\text{PLQ}}(y \mid \zeta)$ and $\zeta_{\text{PG}} = \text{argmax}_\zeta \, \bar{p}_{\text{PG}}(y \mid \zeta)$, Table 3.1 shows that in fact $\text{KL}[\, \bar{p}_{\text{PG}}(\beta \mid y, \zeta_{\text{PLQ}}) \,\|\, p(\beta \mid y) \,]$ is smaller than $\text{KL}[\, \bar{p}_{\text{PG}}(\beta \mid y, \zeta_{\text{PG}}) \,\|\, p(\beta \mid y) \,]$ in the example considered, while analogous relative improvements are observed also in terms of posterior moments. This constitutes a

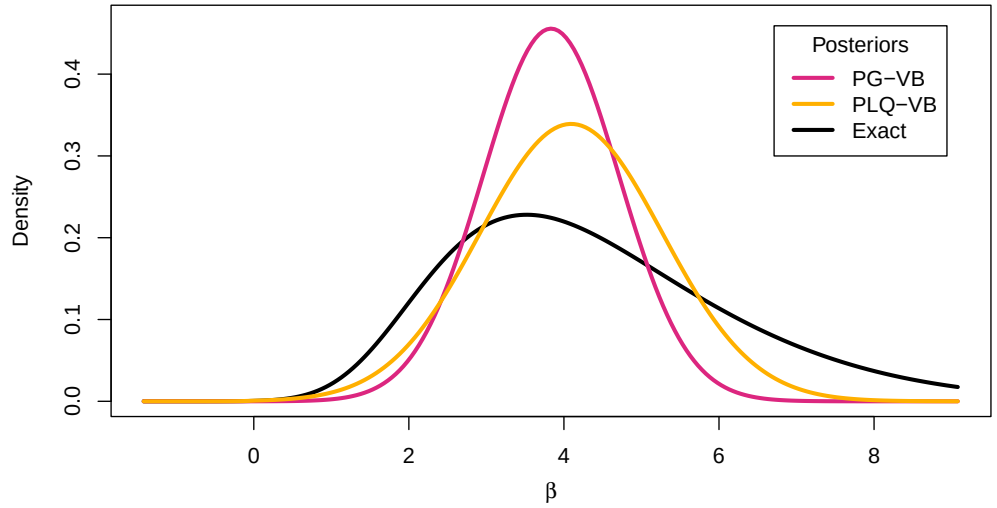

Figure 3.1: The strong skewness of the exact posterior leads to a poor performance of the PG MF-VB distribution, which not only underestimates the variance but also over-shrinks the posterior mean. The proposed PLQ variational approximate leads to an improvement in both directions.

promising result even for the hybrid strategy proposed in Section 3.3.3, which would in fact return a Gaussian approximate posterior virtually indistinguishable from $\bar{p}_{\text{PG}}(\beta \mid y, \zeta_{\text{PLQ}})$.

Table 3.1: The monotonicity of the ELBOs given by different approximations follows from equation (3.10). Although this does not offer guarantees on the posterior moments or the KL from the exact posteriors, both the PLQ and the quadratic approximation evaluated at ζ_{PLQ} improve over the standard MF-VB in these directions. This motivates further investigations on the performance in higher dimensions of the hybrid strategy from Section 3.3.3.

Method	Posterior $\bar{p}(\beta \mid y)$	Marginal Likelihood	Mean	Variance	$KL[\bar{p}(\beta \mid y) \Vert p(\beta \mid y)]$
Exact	$p(\beta \mid y)$	0.2252	4.4144	3.6615	0
PLQ $\mid \zeta_{\text{PLQ}}$	$\bar{p}_{\text{PLQ}}(\beta \mid y, \zeta_{\text{PLQ}})$	0.1396	4.0942	1.3816	0.15497
PLQ $\mid \zeta_{\text{PG}}$	$\bar{p}_{\text{PLQ}}(\beta \mid y, \zeta_{\text{PG}})$	0.1367	3.7822	1.1673	0.20160
PG $\mid \zeta_{\text{PG}}$	$\bar{p}_{\text{PG}}(\beta \mid y, \zeta_{\text{PG}})$	0.1103	3.8335	0.7667	0.33318
PG $\mid \zeta_{\text{PLQ}}$	$\bar{p}_{\text{PG}}(\beta \mid y, \zeta_{\text{PLQ}})$	0.1074	4.2252	0.8451	0.32684

3.5 Discussion

In the current Chapter, we proposed a refined variational approximation for the exact posterior distribution that arises in logistic regression models under Gaussian prior for the coefficients. We tackled the problem by leveraging on a novel piece-wise quadratic tangent bound for logistic log-likelihood, introduced in the previous Chapter, that dominates over the tightest quadratic minorizer by Jaakkola & Jordan (2000). As the use of the latter in variational inference coincides with a proper MF-VB under the Pólya-Gamma data augmentation (Polson et al., 2012; Durante & Rigon, 2019), our contribution fits within the broader endeavor in the literature to improve over the poor performances of mean-field variational approximation in high-dimensional scenarios (Hoffman & Blei, 2015; Guo et al., 2016; Miller et al., 2017; Fasano et al., 2022). Despite more elaborate than purely Gaussian approximations, the proposed variational Bayes scheme allows for efficient and accurate calculations via a well-known scale mixture representation for Laplace random variables (Park & Casella, 2008; Kozumi & Kobayashi, 2011). The same strategy is not applicable for alternative contributions in the literature, improving over the bound by Jaakkola & Jordan (2000) via piece-wise quadratic surrogates (Khan et al., 2010; Marlin et al., 2011; Ermis & Bouchard, 2014). At the same time, the PLQ bound can be used to construct a hybrid variational approximation, that retains the tractability and analytical expression of the PG MF-VB posterior but optimizes the associated variational parameters by considering a modified objective function (Neal & Hinton, 1998; Gunawardana & Byrne, 2005; Parizi et al., 2019). This translates to a simple perturbation of the mean-field fixed-point iteration scheme, that leads to substantial empirical improvement without affecting computational complexity. In particular, the resulting approximation substantially corrects the over-shrinking of the approximate posterior mean, and to a lesser extent also the concurrent variance underestimation. As such, it bears some interesting qualitative similarities with the linear response variational Bayes correction method by Giordano et al. (2015). In fact, as the latter crucially assumes reliable estimates for the posterior means, a combination with our proposed methodology is most promising.

Finally, we plan to further investigate the novel piece-wise quadratic bound, in an effort to establish a formal connection with a full-fledged mean-field variational Bayes routine. In practice, this would consist in a reverse-engineering process to construct or recognize a data augmentation scheme for logistic regression that induces the aforementioned refined bound (Gramacy & Polson, 2012; Polson & Scott, 2013; Bhadra et al., 2020), endowing the proposed methodology with a full probabilistic interpretation. Furthermore, the availability of such a hierarchical representation would potentially allow to greatly simplify calculations, for instance by exploiting the scale mixture representation of Laplace contributions at the level of the latent variables.

Discussion

In the present thesis, we developed novel methodological and computational contributions for statistical inference in high-dimensional regression settings. Tackling the problem mostly from a Bayesian perspective, we focused on some of the most popular regression formulations dealing with discrete-valued observations, both binary and categorical, building on probit and logit link functions. The discrete nature of the observations poses various challenges to Bayesian inference in several respects, including hindered analytical calculations and mixing issues within data augmentation-based sampling schemes.

In the case of probit model, conjugacy results for the regression coefficients under SUN priors were proven only recently by Durante (2019). Such original results were then rapidly extended to a number of constructions generalizing standard probit regression. In Chapter 1, we reviewed and unified such recent advances, developing a unifying general framework that encompasses a broad class of statistical formulations, crucially all relying on the partial discretization of a set of latent linear Gaussian regressions. Among others, this accounts for standard, multinomial and multivariate probit models, tobit models and several related extensions, other than corresponding dynamic, skewed and non-linear formulations, including Gaussian processes (Fasano & Durante, 2022; Fasano et al., 2021; Cao et al., 2022; Benavoli et al., 2020). Furthermore, the generic form of the likelihood function associated with the developed framework implies conjugacy under SUN prior for all the accounted regression models. This result leads to new theoretical insights, benefiting from several useful properties of the SUN posterior, as well as to the development of a unifying computational framework for a broad class of models potentially involving discrete-valued observations. This chiefly includes i.i.d. sampling from exact unified skew–normal posterior, taking advantage of an additive stochastic representation admitted by SUN random variables. As this involves sampling from a multivariate truncated normal distribution, we review a rich literature dealing with such a task (e.g., Botev, 2017; Gessner et al., 2020), which is indeed a longstanding and recurrent problem in various statistical constructions. In particular, the property of the SUN posterior allows to easily handle i.i.d. sampling in large-p-small-n settings, where alternative MCMC schemes encounter different limitations. Nonetheless, our results highlight the importance of further research efforts to develop an efficient gold-standard approach to sample from multivariate truncated Gaussian distributions and to evaluate multivariate normal distribution functions. To deal with settings where the number of discrete observations exceeds a few hundred, we focus also on de-

terministic approximation approaches to posterior inference. In particular, we extend a recently derived partially-factorized variational Bayes scheme (Fasano & Rebaudo, 2021) to all models encompassed by the developed unified framework. Concurrently, we develop a novel scalable implementation of expectation-propagation, characterized by a linear cost per iteration in the number of covariates, unachieved by any alternative implementation available in the literature (Chopin & Ridgway, 2017).
Through our work, we considered the covariance matrices $\bar{\boldsymbol{\Sigma}}_1$ and $\bar{\boldsymbol{\Sigma}}_0$ to be fixed, while we concentrated our effort on inference for the regression parameter $\boldsymbol{\beta}$. Nonetheless, we note that it is possible to produce empirical Bayes estimates of such quantities via numerical maximization of the marginal likelihood $p(\mathbf{y})$ in (1.16), eventually relying on the corresponding accurate approximate version that arises from EP. Indeed, such matrices are often parameterized by a low–dimensional vector of parameters, as exemplified in Sections 1.2.1–1.2.4, which makes the direct maximization of $p(\mathbf{y})$ a feasible approach. Alternatively, full Bayesian inference for the same covariance matrices $\bar{\boldsymbol{\Sigma}}_1$ and $\bar{\boldsymbol{\Sigma}}_0$, once assigned a suitable prior distribution, might proceed by still taking advantage from the availability of a closed–form expression (1.16) for $p(\mathbf{y})$, coupled with i.i.d. sampling schemes from $(\boldsymbol{\beta} \mid \mathbf{y})$. For instance, this would allow to derive collapsed Metropolis–Hastings routines to sample from the posteriors of $\bar{\boldsymbol{\Sigma}}_1$ and $\bar{\boldsymbol{\Sigma}}_0$ after integrating out $\boldsymbol{\beta}$ analytically, benefiting from improved mixing over MCMC schemes based on full–conditional distributions both for $\boldsymbol{\beta}$ and a set of augmented data (e.g., Park & Van Dyk, 2009; Chan & Jeliazkov, 2009).
Concerning the choice of the hyperparameters within the SUN prior, we note that it might not be straightforward to elicit domain knowledge information while maintaining the most general SUN representation, particularly dealing with the matrices $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$ and for large values of $\bar{n}$. Although the role of such quantities within the generative process that leads to the SUN density might help prior specification, this matter surely deserves further investigation. Nonetheless, since multivariate normal distributions are special cases of SUNs, all the uninformative, weakly informative and informative priors relying on Gaussians (e.g., Zellner, 1986; Gelman et al., 2008; Chopin & Ridgway, 2017) can be employed by letting $\bar{n} = 0$ and suitably specifying $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$. Conversely, it remains possible to readily incorporate prior information on the skewness for $\boldsymbol{\beta}$ by eliciting simpler structures. For instance, a convenient choice might be to place independent univariate skew–normals priors for each coefficient β_j, $j = 1, \dots, \bar{p}$, which allow to regulate prior skewness via a single and interpretable parameter for each coefficient. Finally, it would be also worth to include hyperpriors for the scale parameters of the Gaussian or, more generally, SUN prior, which would yield scale–mixture representations that induce shrinkage in high dimensions (Carvalho et al., 2010). Since most of these constructions rely on conditionally Gaussian priors, the results in the present review may be useful to obtain improved theoretical and practical performance in state–of–the–art implementations of the models in Section 1.2 under sparse settings and more general classes of priors.

In Chapter 2 and 3 we still focused on high-dimensional regression with discrete-valued

data, yet turning our attention to binary regression under logit link. We first consider maximum likelihood estimation inference for logistic regression models, both with and without lasso and ridge regularizations. Given the lack of closed-form expressions even in the absence of such penalties, ML estimation typically proceeds by via sequential optimization schemes, which entail the specification of a family of surrogate targets locally approximating the logistic log-likelihood. As a naive Taylor expansion is known to be prone to unreliable and unstable behaviors, especially dealing with extreme values of choice probabilities, we focus on minorize-maximize and expectation-maximization schemes, which benefit from solid convergence guarantees (Böhning & Lindsay, 1988; McLachlan & Krishnan, 1996). We first consider a tractable quadratic minorizer, that has been recently shown to arise from Pólya–Gamma data augmentation scheme (Durante & Rigon, 2019). Over than showing its optimality among quadratic tangent lower bounds, intended as pointwise tightness to the true target function, we underline the benefit of its use in combination with ℓ_1 penalties, which results in faster convergence of the associated coordinate-wise optimization schemes. Furthermore, we derive a novel piece-wise quadratic tangent minorizer that dominates over the PG bound while allowing greater tractability than alternative piece-wise quadratic surrogates available in the literature (Marlin et al., 2011). The tighter approximation given by the new bound is shown to lead to additional benefit in terms of speed of convergence.

Furthermore, in Chapter 3 we take advantage of such novel lower bound to construct a refined variational approximation of the logit posterior under Gaussian prior. Indeed, several contributions tackling Bayesian logistic regression have concentrated on the development of data augmentation schemes resulting in plain Gibbs sampling strategies, from the well-established Pólya–Gamma DA by Polson et al. (2012), to recent contributions such as the ultimate Pólya–Gamma sampler by Zens et al. (2020). Nonetheless, similarly to the probit case discussed in Chapter 1, these sampling schemes often incur in sever limitations in large-p-small-n scenarios. While in some cases it is possible to introduce further latent working parameters to mitigate such issues by re-scaling and re-centering the augmented data (Zens et al., 2020), variational Bayes routines have become an increasingly popular alternative for approximate posterior inference. This typically amounts to approximating the exact posterior with the closest member within a pre-specified family of distributions. Among several alternatives, the combination of forward KL minimization with mean-field independence assumptions owes its widespread use to the resulting simplicity and great efficiency in conditionally conjugate exponential family models (Blei et al., 2017). However, the resulting approximations often exhibit reduced accuracy in skewed or high-dimensions settings, motivating the development of more refined approximations schemes. To pursue this goal in logistic regression models, we employ a tangent minorizaton perspective, under which the MF-VB for PG DA was originally derived, and combine it with the novel piece-wise linear quadratic bound introduced in Chapter 2. This is shown to lead to more accurate evidence lower bounds, at the same time improving also the estimates of poste-

rior moments. Despite the reduced numerical tractability, the cost of the procedure can be made linear in the number of covariates by exploiting a renowned scale mixture representation for Laplace random variable, thus maintaining a beneficial tractability-accuracy trade-off in high-dimensional scenarios.
Both the proposed methodologies based on the novel piece-wise linear-quadratic approximation still present open challenges and compelling research venues. For what concerned penalized maximum likelihood estimation, this amount foremost to developing novel strategies for the efficient and accurate optimization of the non-smooth lower bound. One possible way to achieve this result would be to leverage on the tight connection with generalized lasso penalties, extending the path-wise optimization routines tailored for such regularizations (Tibshirani & Taylor, 2011; Arnold & Tibshirani, 2016) Alternatively, it would be worth pursuing other relaxations of the associated optimization problem, such as the semi-smooth coordinate descent presented in Section 2.4.3, or the combination with techniques from the stochastic optimization literature, which has become increasingly popular in recent years. Conversely, the variational Bayes approximation from Chapter 3 certainly deserves further empirical and theoretical investigation. The latter might proceed by extending the analysis by (Ghosh et al., 2022), which provided guarantees on the statistical optimality of the variational approximation by Jaakkola & Jordan (2000) from a purely tangent minorization perspective, instead of exploiting the full probabilistic interpretation provided by Durante & Rigon (2019). However, their result focuses on asymptotic regimes in the number of observations, whereas the proposed bound is expected to be superior in large-p-small-n scenarios. A the same time, while the hybrid strategy via generalized alternating minimization proposed in Section 3.3.3 required more careful analysis and detailed implementation, it would be worth investigating the combination of the proposed methodology with the linear response variational correction by Giordano et al. (2015). Finally, more effort should be concentrated on formally recasting the novel piece-wise quadratic bound within a full-fledged variational Bayes routine. This might be possible via a sort of reverse-engineering process, eliciting a data augmentation scheme for logistic regression that induces the aforementioned refined bound (Gramacy & Polson, 2012; Polson & Scott, 2013; Bhadra et al., 2020). Other than endowing the proposed methodology with a full probabilistic interpretation, the availability of such a hierarchical representation would potentially allow to greatly simplify calculations, for instance by exploiting the scale mixture representation of Laplace contributions at the level of the latent variables.

Appendix

A.1 Scalable EP implementation

A.1.1 Naive implementation of EP

In the next two Sections, we detail the novel scalable implementation for EP in the setting of equation (1.26) from Chapter 1. For ease of exposition, we first report in Algorithm 1 the pseudocode corresponding to the naive implementation of the procedure described in Section 1.4.3. To do so, we re-formulate the moments in equation (1.29) as to highlight a specific algebraic structure, that we leverage on in the following Sections to derive efficient implementations of the EP updates. Specifically, in accordance with equation (1.13), the first to moments of the hybrid distribution $p_h^{(t_c)}(\boldsymbol{\beta} \mid \mathbf{y})$ read

$$\begin{aligned} \mathbb{E}_h^{(t_c)}[\boldsymbol{\beta} \mid \mathbf{y}] &= \boldsymbol{\xi}_c + \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{s}_c^{-1} \boldsymbol{\psi}_c, \\ \text{var}_h^{(t_c)}[\boldsymbol{\beta} \mid \mathbf{y}] &= \boldsymbol{\Omega}_c + \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{s}_c^{-1} \boldsymbol{\Pi}_c \mathbf{s}_c^{-1} \bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c, \end{aligned} \tag{14}$$

where

$$\begin{aligned} \boldsymbol{\psi}_c &= \boldsymbol{\Gamma}_c^{-1} \mathbb{E}[\mathbf{U}_{1c}], \\ \boldsymbol{\Pi}_c &= \boldsymbol{\Psi}_c - \boldsymbol{\psi}_c \boldsymbol{\psi}_c^\intercal = \boldsymbol{\Gamma}_c^{-1} \big(\text{var}[\mathbf{U}_{1c}] - \boldsymbol{\Gamma}_c \big) \boldsymbol{\Gamma}_c^{-1}, \end{aligned} \tag{15}$$

with $\mathbf{U}_{1c} \sim \text{TN}_{\bar{n}_c}(-\boldsymbol{\gamma}_c; \mathbf{0}, \boldsymbol{\Gamma}_c)$ as before.

In the presentation of EP from Section 1.4.3, we omitted the normalizing constants of the hybrid and approximate distributions, for ease of exposition. We report here the details on how to calculate and update these constants for completeness, since they are needed to assess the approximation of the marginal likelihood provided by EP. In particular, the normalized global approximation reads

$$q_{\text{EP}}(\boldsymbol{\beta}) = \prod_{c=0}^{C} \frac{1}{Z_c} q_c(\boldsymbol{\beta}) = \exp\left\{ -\frac{1}{2} \boldsymbol{\beta}^\intercal \mathbf{Q}_{\text{EP}} \boldsymbol{\beta} + \boldsymbol{\beta}^\intercal \mathbf{r}_{\text{EP}} - \Upsilon(\mathbf{r}_{\text{EP}}, \mathbf{Q}_{\text{EP}}) \right\}$$

where $\Upsilon(\mathbf{r}_{\text{EP}}, \mathbf{Q}_{\text{EP}}) = \frac{1}{2} \mathbf{r}_{\text{EP}}^\intercal \mathbf{Q}_{\text{EP}}^{-1} \mathbf{r}_{\text{EP}} + \frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{Q}_{\text{EP}}|$ accounts for the normalization of the Gaussian distribution in its natural parametrization. As before, at step t_c we refine the c-th site indirectly, as we obtain an updated global Gaussian approximation $q_{\text{EP}}^{(t_c)}(\boldsymbol{\beta})$ my

Algorithm 1: EP algorithm for the factorized likelihood in equation (1.26)

$\mathbf{Q}_{\text{EP}} = \mathbf{Q}_0 = \boldsymbol{\Omega}_{\text{post}}^{-1} = \boldsymbol{\Omega}^{-1} + \bar{\mathbf{X}}_1^\intercal \bar{\boldsymbol{\Sigma}}_1^{-1} \bar{\mathbf{X}}_1$

$\mathbf{r}_{\text{EP}} = \mathbf{r}_0 = \boldsymbol{\Omega}_{\text{post}}^{-1} \boldsymbol{\xi}_{\text{post}} = \boldsymbol{\Omega}^{-1} \boldsymbol{\xi} + \bar{\mathbf{X}}_1^\intercal \bar{\boldsymbol{\Sigma}}_1^{-1} \bar{\mathbf{y}}_1$

for $c = 1, \ldots, C$ **do**

  $\mathbf{Q}_c = \mathbf{0}_{p \times p}$; $\mathbf{r}_c = \mathbf{0}_p$; $\log(1/Z_c) = 0$

end

for t *from* 1 *until convergence* **do**

  for c *from* C **do**

    ..... Cavity distribution

    $\mathbf{Q}_{-c} = \mathbf{Q}_{\text{EP}} - \mathbf{Q}_c$

    $\mathbf{r}_{-c} = \mathbf{r}_{\text{EP}} - \mathbf{r}_c$

    ..... Hybrid distribution

    $\boldsymbol{\Omega}_c = \mathbf{Q}_{-c}^{-1}$

    $\boldsymbol{\xi}_c = \boldsymbol{\Omega}_c \, \mathbf{r}_{-c}$

    $\mathbf{s}_c = ([\bar{\boldsymbol{\Sigma}}_{0[c,c]} + \bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal] \odot \mathbf{I}_{\bar{n}_c})^{1/2}$

    $\boldsymbol{\gamma}_c = \mathbf{s}_c^{-1} (\bar{\mathbf{y}}_{0[c]} + \bar{\mathbf{X}}_{0[c]} \boldsymbol{\xi}_c)$

    $\boldsymbol{\Gamma}_c = \mathbf{s}_c^{-1} (\bar{\boldsymbol{\Sigma}}_{0[c,c]} + \bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal) \mathbf{s}_c^{-1}$

    ..... Moments evaluation

    $\boldsymbol{\psi}_c = \boldsymbol{\Gamma}_c^{-1} \mathbb{E}[\mathbf{U}_{1c}]$

    $\boldsymbol{\Pi}_c = \boldsymbol{\Gamma}_c^{-1} \big(\text{var}[\mathbf{U}_{1c}] - \boldsymbol{\Gamma}_c \big) \boldsymbol{\Gamma}_c^{-1}$

    ..... Global approximation

    $\mathbf{Q}_{\text{EP}} = \big(\boldsymbol{\Omega}_c + (\boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal) \mathbf{s}_c^{-1} \boldsymbol{\Pi}_c \mathbf{s}_c^{-1} (\bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c) \big)^{-1}$

    $\mathbf{r}_{\text{EP}} = \mathbf{Q}_{\text{EP}} \big(\boldsymbol{\xi}_c + (\boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal) \mathbf{s}_c^{-1} \boldsymbol{\psi}_c \big)$

    ..... c-th site approximation

    $\mathbf{Q}_c = \mathbf{Q}_{\text{EP}} - \mathbf{Q}_{-c}$

    $\mathbf{r}_c = \mathbf{r}_{\text{EP}} - \mathbf{r}_{-c}$

    $\log(1/Z_c) = \log \Phi_{n_c}(\boldsymbol{\gamma}_c, \boldsymbol{\Gamma}_c) + \frac{1}{2} \log |\mathbf{Q}_{\text{EP}}| - \frac{1}{2} \log |\mathbf{Q}_{-c}| +$
    $\quad - \frac{1}{2} \mathbf{r}_{\text{EP}}^\intercal \mathbf{Q}_{\text{EP}}^{-1} \mathbf{r}_{\text{EP}} + \frac{1}{2} \mathbf{r}_{-c}^\intercal \mathbf{Q}_{-c}^{-1} \mathbf{r}_{-c}$

  end

end

$\boldsymbol{\Omega}_{\text{EP}} = \mathbf{Q}_{\text{EP}}^{-1}$

$\boldsymbol{\xi}_{\text{EP}} = \boldsymbol{\Omega}_{\text{EP}} \, \mathbf{r}_{\text{EP}}$

$\log q(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1) = \frac{1}{2} \mathbf{r}^\intercal \boldsymbol{\xi}_{\text{EP}} - \frac{1}{2} \mathbf{r}_0^\intercal \mathbf{Q}_0^{-1} \mathbf{r}_0 - \frac{1}{2} \log |\mathbf{Q}_{\text{EP}}| + \frac{1}{2} \log |\mathbf{Q}_0| + \sum_{c=1}^{C} \log(1/Z_c)$

Result: ($\boldsymbol{\Omega}_{\text{EP}}, \boldsymbol{\xi}_{\text{EP}}, \log q(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1)$)

moment matching with the hybrid $p_h^{(t_c)}(\boldsymbol{\beta})$, including its zeroth order moment

$$Z_h^{(t_c)} = \int l_c(\boldsymbol{\beta}) \prod_{c'=1}^{C} q_{c'}^{(t_c)}(\boldsymbol{\beta})\, d\boldsymbol{\beta} = \Phi_{\bar{n}_c}(\boldsymbol{\gamma}_c; \boldsymbol{\Gamma}_c).$$

Then, we reconstruct the updated c-th site $q_c^{(t_c)}(\boldsymbol{\beta})$ so that its combination with the remaining unaltered ones gives exactly $q_{\text{EP}}^{(t_c)}(\boldsymbol{\beta})$

$$\frac{1}{Z_c^{(t_c)}} q_c^{(t_c)}(\boldsymbol{\beta}) = Z_h^{(t_c)} \frac{q_{\text{EP}}^{(t_c)}(\boldsymbol{\beta})}{q_{-c}^{(t_c)}(\boldsymbol{\beta})},$$

where in the equation above both the cavity and the global approximation are intended as properly normalized distributions. Accordingly

$$\log 1/Z_c^{(t_c)} = \log Z_h^{(t_c)} - \Upsilon(\mathbf{r}_{\text{EP}}^{(t_c)}, \mathbf{Q}_{\text{EP}}^{(t_c)}) + \Upsilon(\mathbf{r}_{-c}^{(t_c)}, \mathbf{Q}_{-c}^{(t_c)}).$$

Likewise, after the convergence of the iterations, the global approximation of the marginal log-likelihood $\log p(\mathbf{y}) = \log p(\bar{\mathbf{y}}_1) + \log p(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1)$ is obtained by replacing $p(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1)$ with its EP equivalent

$$q_{\text{EP}}(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1) = \int \prod_{c=0}^{C} \frac{1}{Z_c} q_c(\boldsymbol{\beta})\, d\boldsymbol{\beta} = \frac{\exp\{\Upsilon(\mathbf{r}_{\text{EP}}, \mathbf{Q}_{\text{EP}})\}}{\prod_{c=0}^{C} Z_c}\,, \tag{16}$$

where $Z_0 = \exp\{\Upsilon(\mathbf{r}_0, \mathbf{Q}_0)\}$.

A.1.2 Scalable Gaussian EP for probit and tobit regression

In this Section, we present an efficient implementation of expectation–propagation for a specific set of models among the ones embedded in equation (1.1). Specifically, we focus here on likelihoods of the form

$$p(\mathbf{y} \mid \boldsymbol{\beta}) = \phi_{\bar{n}_1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{X}}_1\boldsymbol{\beta}; \sigma^2 \mathbf{I}_{n_1}) \prod_{i=1}^{\bar{n}_0} \Phi(\bar{\mathbf{y}}_{0[i]} + \bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\beta}),$$

which in particular accounts both for standard tobit and probit regression. Under a Gaussian prior $p(\boldsymbol{\beta}) = \phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega})$, the factorized target distribution becomes

$$p(\boldsymbol{\beta} \mid \mathbf{y}) = \frac{1}{p(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1)} \phi_{\bar{p}}(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}; \boldsymbol{\Omega}_{\text{post}}) \prod_{i=1}^{\bar{n}_0} \Phi(\bar{\mathbf{y}}_{0[i]} + \bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\beta})$$

Accordingly, we recognize in the hybrid distribution

$$p_h(\boldsymbol{\beta}) \propto \Phi(\bar{\mathbf{y}}_{0[i]} + \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\beta}) \exp\left\{ -\frac{1}{2}\boldsymbol{\beta}^\intercal \mathbf{Q}_{-i}\boldsymbol{\beta} + \boldsymbol{\beta}^\intercal \mathbf{r}_{-i} \right\}$$

the kernel of a multivariate extended skew normal $\mathrm{SN}_{\bar{p}}(\boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\alpha}_i, \gamma_i)$ random variable, with

$$\begin{aligned} \boldsymbol{\Omega}_i &= \mathbf{Q}_{-i}^{-1} = \boldsymbol{\omega}_i \bar{\boldsymbol{\Omega}}_i \boldsymbol{\omega}_i & \boldsymbol{\alpha}_i^\intercal &= \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\omega}_i \\ \boldsymbol{\xi}_i &= \boldsymbol{\Omega}_i \mathbf{r}_{-i} & \gamma_i &= (1 + \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]})^{-1/2} (\bar{\mathbf{y}}_{0[i]} + \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\xi}_i) \,, \end{aligned}$$

Its normalization constant is $Z_h = \Phi(\gamma_i)$, while the first to moments read

$$\begin{aligned} \mathbb{E}_h[\boldsymbol{\beta} \mid \mathbf{y}] &= \boldsymbol{\xi}_i + \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]}\, \psi_1(\gamma_i) s_i \\ \mathrm{var}_h[\boldsymbol{\beta} \mid \mathbf{y}] &= \boldsymbol{\Omega}_i + \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]}\, \psi_2(\gamma_i) s_i^2\, \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \end{aligned}$$

where $s_i = (1 + \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]})^{-1/2}$, $\psi_1(x) = \phi(x)/\Phi(x)$ and $\psi_2(x) = -\psi_1(x)^2 - x\psi_1(x)$. As before, we update the global approximation parameters $\mathbf{Q}_{\mathrm{EP}}^{(t_i)}$ and $\mathbf{r}_{\mathrm{EP}}^{(t_i)}$ by moment matching with the hybrid, and modify accordingly the parameters of i-th approximate site $\mathbf{Q}_i^{(t_i)}$ and $\mathbf{r}_i^{(t_i)}$. For ease of notation, in the remaining of this appendix we replace the superscript "(t_i)" on the updated parameters with "new", and drop the superscript "(t_{i-1})" identifying the outcome of the previous EP update. The key step for simplifying Algorithm 1, adapted to the one-dimensional sites under consideration, is to employ the Woodbury identity for calculating the matrix inverse $(\mathrm{var}_h[\boldsymbol{\beta}])^{-1}$. Specifically, we rewrite the updated site precision matrix as

$$\begin{aligned} \mathbf{Q}_i^{\mathrm{new}} &= \big(\mathrm{var}_h[\boldsymbol{\beta} \mid \mathbf{y}]\big)^{-1} - \mathbf{Q}_{-i} = \big(\boldsymbol{\Omega}_i + \psi_2(\gamma_i) s_i^2 (\boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]})(\boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]})^\intercal\big)^{-1} - \mathbf{Q}_{-i} \\ &= \boldsymbol{\Omega}_i^{-1} - \psi_2(\gamma_i) s_i^2 (1 + \psi_2(\gamma_i) s_i^2 \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]})^{-1} \bar{\mathbf{x}}_{0[i]} \bar{\mathbf{x}}_{0[i]}^\intercal - \boldsymbol{\Omega}_i^{-1} \\ &= \left[-(\psi_2(\gamma_i)^{-1} s_i^{-2} + \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]})^{-1} \right] \bar{\mathbf{x}}_{0[i]} \bar{\mathbf{x}}_{0[i]}^\intercal \\ &= \left[-\frac{\psi_2(\gamma_i)}{1 + \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]} + \psi_2(\gamma_i) \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]}} \right] \bar{\mathbf{x}}_{0[i]} \bar{\mathbf{x}}_{0[i]}^\intercal = k_i^{\mathrm{new}} \bar{\mathbf{x}}_{0[i]} \bar{\mathbf{x}}_{0[i]}^\intercal \end{aligned}$$

with $k_i^{\mathrm{new}} = -\dfrac{\psi_2(\gamma_i)}{1 + \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]} + \psi_2(\gamma_i) \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]}}$. Moreover, the corresponding new site location vector is obtained as

$$\begin{aligned} \mathbf{r}_i^{\mathrm{new}} &= \mathbf{Q}_{\mathrm{EP}}^{\mathrm{new}} \mathbb{E}_h[\boldsymbol{\beta}] - \mathbf{r}_{-i} = \mathbf{Q}_{-i} \mathbb{E}_h[\boldsymbol{\beta}] + \mathbf{Q}_i^{\mathrm{new}} \mathbb{E}_h[\boldsymbol{\beta}] - \mathbf{r}_{-i} \\ &= \mathbf{Q}_{-i} \mathbf{Q}_{-i}^{-1} \mathbf{r}_{-i} + \psi_1(\gamma_i) s_i \bar{\mathbf{x}}_{0[i]} + \mathbf{Q}_i^{\mathrm{new}} \mathbb{E}_h[\boldsymbol{\beta}] - \mathbf{r}_{-i} \\ &= \psi_1(\gamma_i) s_i \mathbf{Q}_{-i} \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]} + \mathbf{Q}_i^{\mathrm{new}} \mathbb{E}_h[\boldsymbol{\beta}] \\ &= \psi_1(\gamma_i) s_i \bar{\mathbf{x}}_{0[i]} + k_i^{\mathrm{new}} \bar{\mathbf{x}}_{0[i]} \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \mathbf{r}_{-i} + k_i^{\mathrm{new}} \psi_1(\gamma_i) s_i \bar{\mathbf{x}}_{0[i]} \bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]} \\ &= \big[\psi_1(\gamma_i) s_i + k_i^{\mathrm{new}} (\boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]})^\intercal \mathbf{r}_{-i} + k_i^{\mathrm{new}} \psi_1(\gamma_i) s_i (\bar{\mathbf{x}}_{0[i]}^\intercal \boldsymbol{\Omega}_i \bar{\mathbf{x}}_{0[i]}) \big] \bar{\mathbf{x}}_{0[i]} = m_i^{\mathrm{new}} \bar{\mathbf{x}}_{0[i]} \end{aligned}$$

where $m_i^{\text{new}} = \psi_1(\gamma_i)s_i + k_i^{\text{new}}(\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]})^\intercal\mathbf{r}_{-i} + k_i^{(t_i)}\psi_1(\gamma_i)s_i\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}$. This suggests that we can implement EP by saving only the set of scalars $\{k_i\}_{i=1}^{\bar{n}_0}$ and $\{m_i\}_{i=1}^{\bar{n}_0}$, instead of the full matrices $\{\mathbf{Q}_i\}_{i=1}^{\bar{n}_0}$ and vectors $\{\mathbf{r}_i\}_{i=1}^{\bar{n}_0}$, after having initialized the former to some starting values. In practice, we start the algorithm by setting them to zero, which corresponds to initializing the global approximation to the prior distribution. Afterward, both $\bar{p}\times\bar{p}$ matrix inversions involved in each update can be avoided, since

$$\begin{aligned}\boldsymbol{\Omega}_i = \mathbf{Q}_{-i}^{-1} &= \left(\mathbf{Q}_{\text{EP}} - k_i\bar{\mathbf{x}}_{0[i]}\bar{\mathbf{x}}_{0[i]}^\intercal\right)^{-1} \\ &= \left(\mathbf{Q}_{\text{EP}}\right)^{-1} + \frac{k_i}{1 - k_i\bar{\mathbf{x}}_{0[i]}^\intercal\left(\mathbf{Q}_{\text{EP}}\right)^{-1}\bar{\mathbf{x}}_{0[i]}}\left(\mathbf{Q}_{\text{EP}}\right)^{-1}\bar{\mathbf{x}}_{0[i]}\bar{\mathbf{x}}_{0[i]}^\intercal\left(\mathbf{Q}_{\text{EP}}\right)^{-1}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Omega}_{\text{EP}}^{\text{new}} &= \boldsymbol{\Omega}_i + \boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}\,\psi_2(\gamma_i)s_i^2\,\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i \\ \mathbf{Q}_{\text{EP}}^{\text{new}} &= \mathbf{Q}_{-i} + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}\bar{\mathbf{x}}_{0[i]}^\intercal.\end{aligned}$$

Finally, even the update of $\log(1/Z_i^{\text{new}})$ can be simplified noticing that

$$\begin{aligned}&= \frac{1}{2}\log|\mathbf{Q}_{\text{EP}}^{\text{new}}| - \frac{1}{2}\log|\mathbf{Q}_{-i}| \\ &= \frac{1}{2}\log|\mathbf{Q}_{-i} + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}\bar{\mathbf{x}}_{0[i]}^\intercal| - \frac{1}{2}\log|\mathbf{Q}_{-i}| \\ &= \frac{1}{2}\log|\mathbf{Q}_{-i}| + \frac{1}{2}\log\left(1 + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^\intercal\mathbf{Q}_{-i}^{-1}\bar{\mathbf{x}}_{0[i]}\right) - \frac{1}{2}\log|\mathbf{Q}_{-i}| \\ &= \frac{1}{2}\log\left(1 + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}\right)\end{aligned}$$

and

Algorithm 2: Probit & tobit EP - $\mathcal{O}(\bar{p}^2 \cdot \bar{n}_0)$ cost per iteration

$\mathbf{Q}_{\text{EP}} = \mathbf{Q}_0 = (\boldsymbol{\Omega}^{-1} + \sigma^{-2}\bar{\mathbf{X}}_1^\intercal\bar{\mathbf{X}}_1)$
$\mathbf{r}_{\text{EP}} = \mathbf{r}_0 = (\boldsymbol{\Omega}^{-1}\boldsymbol{\xi} + \sigma^{-2}\bar{\mathbf{X}}_1^\intercal\bar{\mathbf{y}}_1)$
for $i = 1, \dots, \bar{n}_0$ **do**
  $k_i = 0$; $m_i = 0$; $\log(1/Z_i) = 0$
end
if $\bar{p} > \bar{n}_1$ **then**
  $\boldsymbol{\Omega}_{\text{EP}} = \boldsymbol{\Omega} - \boldsymbol{\Omega}\bar{\mathbf{X}}_1^\intercal(\sigma^2\mathbf{I}_{\bar{n}_1} + \bar{\mathbf{X}}_1\boldsymbol{\Omega}X_1^\intercal)^{-1}\bar{\mathbf{X}}_1\boldsymbol{\Omega}$
else
  $\boldsymbol{\Omega}_{\text{EP}} = (\mathbf{Q}_0)^{-1}$
end
$\boldsymbol{\xi}_{\text{post}} = \mathbf{Q}_0^{-1}\mathbf{r}_0$
for t *from* 1 *until convergence* **do**
  for i *from* $\bar{n}_0$ **do**
    ..... Cavity distribution
    $\mathbf{Q}_{-i} = \mathbf{Q}_{\text{EP}} - k_i\bar{\mathbf{x}}_{0[i]}\bar{\mathbf{x}}_{0[i]}^\intercal$
    $\mathbf{r}_{-i} = \mathbf{r}_{\text{EP}} - m_i\bar{\mathbf{x}}_{0[i]}$
    ..... Hybrid distribution
    $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}_{\text{EP}} + \dfrac{k_i}{1 - k_i\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_{\text{EP}}\bar{\mathbf{x}}_{0[i]}}(\boldsymbol{\Omega}_{\text{EP}}\bar{\mathbf{x}}_{0[i]})(\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_{\text{EP}})$
    $s_i = (1 + \bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]})^{-1/2}$
    $\gamma_i = s_i\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\,\mathbf{r}_{-i}$
    ..... i-th site approximation
    $k_i = -\psi_2(\gamma_i)/\big(1 + \bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]} + \psi_2(\gamma_i)\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}\big)$
    $m_i = \psi_1(\gamma_i)s_i + k_i(\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i)\mathbf{r}_{-i} + k_i\psi_1(\gamma_i)s_i\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}$
    $\log(1/Z_i) = \log\Phi(\gamma_i) + \dfrac{1}{2}\log\Big(1 + k_i\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}\Big) +$
    $-\dfrac{1}{2}\dfrac{\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}{1 + k_i\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}\left(m_i + \dfrac{\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\mathbf{r}_{-i}}{\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}\right)^2 + \dfrac{1}{2}\dfrac{\Big(\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\mathbf{r}_{-i}\Big)^2}{\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}$
    ..... Global approximation
    $\mathbf{Q}_{\text{EP}} = \mathbf{Q}_{-i} + k_i\bar{\mathbf{x}}_{0[i]}\bar{\mathbf{x}}_{0[i]}^\intercal$
    $\mathbf{r}_{\text{EP}} = \mathbf{r}_{-i} + m_i\bar{\mathbf{x}}_{0[i]}$
    $\boldsymbol{\Omega}_{\text{EP}} = \boldsymbol{\Omega}_i + \boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}\,\psi_2(\gamma_i)s_i^2\,\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\Omega}_i$
  end
end
$\boldsymbol{\xi}_{\text{EP}} = \boldsymbol{\Omega}_{\text{EP}}\,\mathbf{r}_{\text{EP}}$
$\log q(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1) = \frac{1}{2}\mathbf{r}_{\text{EP}}^\intercal\boldsymbol{\xi}_{\text{EP}} - \frac{1}{2}\mathbf{r}_0^\intercal\boldsymbol{\xi}_{\text{post}} - \frac{1}{2}\log|\mathbf{Q}_{\text{EP}}| + \frac{1}{2}\log|\mathbf{Q}_0| + \sum_{i=1}^{\bar{n}_0}\log(1/Z_i)$
Result: ($\boldsymbol{\Omega}_{\text{EP}}, \boldsymbol{\xi}_{\text{EP}}, \log q(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1)$)

$$
\begin{aligned}
&= -\frac{1}{2}\mathbf{r}_{\text{EP}}^{\text{new}\intercal}(\mathbf{Q}_{\text{EP}}^{\text{new}})^{-1}\mathbf{r}_{\text{EP}}^{\text{new}} + \frac{1}{2}\mathbf{r}_{-i}^{\intercal}\boldsymbol{\Omega}_i\mathbf{r}_{-i} \\
&= -\frac{1}{2}(\mathbf{r}_{-i}^{\intercal} + m_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^{\intercal})\left(\boldsymbol{\Omega}_i - \frac{k_i^{\text{new}}}{1 + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}(\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]})(\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i)\right)(\mathbf{r}_{-i} + m_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}) \\
&\quad + \frac{1}{2}\mathbf{r}_{-i}^{\intercal}\boldsymbol{\Omega}_i\mathbf{r}_{-i} \\
&= -\frac{1}{2}\mathbf{r}_{-i}^{\intercal}\boldsymbol{\Omega}_i\mathbf{r}_{-i} - \frac{1}{2}(m_i^{\text{new}})^2\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]} - m_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\mathbf{r}_{-i} \\
&\quad + \frac{1}{2}\frac{k_i^{\text{new}}}{1 + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}(\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\mathbf{r}_{-i})^2 + \frac{1}{2}\frac{k_i^{\text{new}}}{1 + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}(m_i^{\text{new}})^2(\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]})^2 \\
&\quad + \frac{k_i^{\text{new}}}{1 + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}m_i^{\text{new}}(\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]})(\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\mathbf{r}_{-i}) + \frac{1}{2}\mathbf{r}_{-i}^{\intercal}\boldsymbol{\Omega}_i\mathbf{r}_{-i} \\
&= -m_i^{\text{new}}(\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\mathbf{r}_{-i})\frac{1}{1 + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}} - \frac{1}{2}(m_i^{\text{new}})^2(\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]})\frac{1}{1 + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}} \\
&\quad + \frac{1}{2}(\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\mathbf{r}_{-i})^2\frac{k_i^{\text{new}}}{1 + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}} \\
&= -\frac{1}{2}\frac{\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}{1 + k_i^{\text{new}}\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}\left(m_i^{\text{new}} + \frac{\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\mathbf{r}_{-i}}{\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}\right)^2 + \frac{1}{2}\frac{\left(\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\mathbf{r}_{-i}\right)^2}{\bar{\mathbf{x}}_{0[i]}^{\intercal}\boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}}
\end{aligned}
$$

Assembling all the results above, we can construct the alternative EP scheme reported in Algorithm 2. Except for the efficient calculation of $\log(1/Z_i^{\text{new}})$, the core of the EP updates in Algorithm 2 takes the same form as in the implementation of the R package EPGLM, used as a benchmark by Chopin & Ridgway (2017). Notably, there is no direct $\bar{p} \times \bar{p}$ matrix inversion or determinant calculation, reducing the cost per EP iteration to $\mathcal{O}(\bar{p}^2 \cdot \bar{n}_0)$.

Despite the improvement over a naive implementation, like that of Algorithm 1, the quadratic cost in the number of covariates can still make computations impractically slow for high-dimensional datasets. In the remaining part of this Section we show that the same representation can be further exploited to formulate an equivalent EP scheme with $\mathcal{O}(\bar{p} \cdot \bar{n}_0^2)$ cost per iteration. In fact, we can avoid handling explicitly any $p \times p$ matrix by defining the new variables $\boldsymbol{\tau}_i = \mathbf{Q}_{-i}^{-1}\bar{\mathbf{x}}_{0[i]} = \boldsymbol{\Omega}_i\bar{\mathbf{x}}_{0[i]}$ and $\mathbf{u}_i = \mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{x}}_{0[i]}$, for each $i = 1, \dots, n_0$, and working out directly their updates. Coherently with the above formulation

$$
\begin{aligned}
\boldsymbol{\tau}_i = \mathbf{Q}_{-i}^{-1}\bar{\mathbf{x}}_{0[i]} &= (\mathbf{Q}_{\text{EP}} - \mathbf{Q}_i)^{-1}\bar{\mathbf{x}}_{0[i]} \\
&= \mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{x}}_{0[i]} + (1 - k_i\bar{\mathbf{x}}_{0[i]}^{\intercal}\mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{x}}_{0[i]})^{-1}k_i(\mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{x}}_{0[i]})(\mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{x}}_{0[i]})^{\intercal}\bar{\mathbf{x}}_{0[i]} \\
&= \mathbf{u}_i + k_i(1 - k_i\bar{\mathbf{x}}_{0[i]}^{\intercal}\mathbf{u}_i)^{-1}\mathbf{u}_i\mathbf{u}_i^{\intercal}\bar{\mathbf{x}}_{0[i]} = \mathbf{u}_i + k_i(1 - k_i\bar{\mathbf{x}}_{0[i]}^{\intercal}\mathbf{u}_i)^{-1}(\bar{\mathbf{x}}_{0[i]}^{\intercal}\mathbf{u}_i)\mathbf{u}_i \\
&= \left[1 + \frac{k_i\bar{\mathbf{x}}_{0[i]}^{\intercal}\mathbf{u}_i}{1 - k_i\bar{\mathbf{x}}_{0[i]}^{\intercal}\mathbf{u}_i}\right]\mathbf{u}_i = \left[\frac{1}{1 - k_i\bar{\mathbf{x}}_{0[i]}^{\intercal}\mathbf{u}_i}\right]\mathbf{u}_i.
\end{aligned}
$$

In this reformulation of the problem, the update of any specific i-th factor affects also every

other site, contrarily to Algorithm 2, in that it leads to a modification of all the vectors $\{\mathbf{u}_j\}_{j=1}^{\bar{n}_0}$. In fact,

$$\begin{aligned}
\mathbf{u}_j^{\text{new}} &= (\mathbf{Q}_{\text{EP}}^{\text{new}})^{-1}\bar{\mathbf{x}}_{0[j]} = (\mathbf{Q}_{\text{EP}} - \mathbf{Q}_i + \mathbf{Q}_i^{\text{new}})^{-1}\bar{\mathbf{x}}_{0[j]} \\
&= \left(\mathbf{Q}_{\text{EP}}^{\text{new}} + (k_i^{\text{new}} - k_i)\bar{\mathbf{x}}_{0[i]}\bar{\mathbf{x}}_{0[i]}^\intercal\right)^{-1}\bar{\mathbf{x}}_{0[j]} \\
&= \left((\mathbf{Q}_{\text{EP}})^{-1} - \frac{k_i^{\text{new}} - k_i}{1 + (k_i^{\text{new}} - k_i)\bar{\mathbf{x}}_{0[i]}^\intercal(\mathbf{Q}_{\text{EP}})^{-1}\bar{\mathbf{x}}_{0[i]}}(\mathbf{Q}_{\text{EP}})^{-1}\bar{\mathbf{x}}_{0[i]}\bar{\mathbf{x}}_{0[i]}^\intercal(\mathbf{Q}_{\text{EP}})^{-1}\right)\bar{\mathbf{x}}_{0[j]} \\
&= \mathbf{u}_j - \mathbf{u}_i\frac{(k_i^{\text{new}} - k_i)}{1 + (k_i^{\text{new}} - k_i)\bar{\mathbf{x}}_{0[i]}^\intercal\mathbf{u}_i}\bar{\mathbf{x}}_{0[i]}^\intercal\mathbf{u}_j.
\end{aligned}$$

Instead of cycling over j, these updates can be performed in block by defining a $\bar{p} \times \bar{n}_0$ matrix $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{\bar{n}_0})$. Accordingly

$$\mathbf{U}^{\text{new}} = \mathbf{U} - \mathbf{u}_i \frac{(k_i^{\text{new}} - k_i)}{1 + (k_i^{\text{new}} - k_i)\bar{\mathbf{x}}_{0[i]}^\intercal\mathbf{u}_i}\bar{\mathbf{x}}_{0[i]}^\intercal\mathbf{U}\,.$$

This operation will be the most expensive per site update, being of order $\mathcal{O}(\bar{p}\cdot\bar{n}_0)$. Accordingly, each EP iteration will have cost $\mathcal{O}(\bar{p}\cdot\bar{n}_0^2)$. Contrarily to Algorithm 2, once the procedure has reached convergence we still need to calculate the inverse of the global precision matrix $\mathbf{Q}_{\text{EP}}^{-1}$, other than its determinant $|\mathbf{Q}_{\text{EP}}|$ as in all previous versions. Both these calculations can be optimized as well in their cost with respect to $\bar{p}$, starting from the observation that $\mathbf{Q}_{\text{EP}} = \mathbf{Q}_0 + \sum_{i=1}^{\bar{n}_0}\bar{\mathbf{x}}_{0[i]}k_i\bar{\mathbf{x}}_{0[i]}^\intercal = \mathbf{Q}_0 + \bar{\mathbf{X}}_0^\intercal\mathbf{K}\bar{\mathbf{X}}_0$ with $\mathbf{K} = \text{diag}(\{k_i\}_{i=1}^{\bar{n}_0})$. Defining $\mathbf{\Lambda} = (\mathbf{K}^{-1} + \bar{\mathbf{X}}_0\mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal)^{-1}$, so that $\mathbf{Q}_{\text{EP}}^{-1} = \mathbf{Q}_0^{-1} - \mathbf{Q}_0^{-1}\bar{\mathbf{X}}_0^\intercal\mathbf{\Lambda}\bar{\mathbf{X}}_0\mathbf{Q}_0^{-1}$, one has

$$\begin{aligned}
\mathbf{U} = \mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_0^\intercal &= \mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal - \mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal(\mathbf{K}^{-1} + \bar{\mathbf{X}}_0\mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal)^{-1}\bar{\mathbf{X}}_0\mathbf{\Omega} \\
&= \mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal\left(\mathbf{I}_{\bar{n}_0} - (\mathbf{I}_{\bar{n}_0} + \mathbf{K}\bar{\mathbf{X}}_0\mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal)^{-1}\mathbf{K}\bar{\mathbf{X}}_0\mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal\right) \\
&= \mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal(\mathbf{I}_{\bar{n}_0} + \mathbf{K}\bar{\mathbf{X}}_0\mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal)^{-1} \\
&= \mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal(\mathbf{K}^{-1} + \bar{\mathbf{X}}_0\mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal)^{-1}\mathbf{K}^{-1} = \mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal\mathbf{\Lambda}\mathbf{K}^{-1},
\end{aligned}$$

and thus, recalling that $\mathbf{U}_0 = \mathbf{\Omega}\bar{\mathbf{X}}_0^\intercal = \mathbf{Q}_0^{-1}\bar{\mathbf{X}}_0^\intercal$

$$\mathbf{\Omega}_{\text{EP}} = \mathbf{Q}_{\text{EP}}^{-1} = \mathbf{\Omega} - \mathbf{U}\mathbf{K}\mathbf{U}_0^\intercal\,,$$

while

$$\begin{aligned}
-\frac{1}{2}\log|\mathbf{Q}_{\text{EP}}| + \frac{1}{2}\log|\mathbf{Q}_0| &= -\frac{1}{2}\log|\mathbf{K}^{-1} + \bar{\mathbf{X}}_0\mathbf{Q}_0^{-1}\bar{\mathbf{X}}_0^\intercal| - \frac{1}{2}\log|\mathbf{K}| \\
&= -\frac{1}{2}\log|\mathbf{I}_{\bar{n}_0} + \mathbf{K}\bar{\mathbf{X}}_0\mathbf{U}_0|\,.
\end{aligned}$$

Similarly, since $\mathbf{r} = \mathbf{r}_0 + \sum_{i=1}^{\bar{n}_0} m_i\bar{\mathbf{x}}_{0[i]} = \bar{\mathbf{X}}_0^\intercal\mathbf{m}$, with $\mathbf{m} = (m_1, m_2, \ldots, m_{\bar{n}_0})^\intercal$, one has

Algorithm 3: Efficient probit & tobit EP - $\mathcal{O}(\bar{p} \cdot \bar{n}_0^2)$ cost per iteration

$\mathbf{Q}_0 = (\boldsymbol{\Omega}^{-1} + \sigma^{-2}\bar{\mathbf{X}}_1^\intercal\bar{\mathbf{X}}_1)$
$\mathbf{r}_0 = (\boldsymbol{\Omega}^{-1}\boldsymbol{\xi} + \sigma^{-2}\bar{\mathbf{X}}_1^\intercal\bar{\mathbf{y}}_1)$
for $i = 1, \ldots, \bar{n}_0$ **do**
  $k_i = 0$; $m_i = 0$; $\log(1/Z_i) = 0$
end
if $\bar{p} > \bar{n}_1$ **then**
  $\boldsymbol{\Omega}_{\text{post}} = \boldsymbol{\Omega} - \boldsymbol{\Omega}\bar{\mathbf{X}}_1^\intercal(\sigma^2\mathbf{I}_{\bar{n}_1} + \bar{\mathbf{X}}_1\boldsymbol{\Omega}X_1^\intercal)^{-1}\bar{\mathbf{X}}_1\boldsymbol{\Omega}$
else
  $\boldsymbol{\Omega}_{\text{post}} = (\mathbf{Q}_0)^{-1}$
end
$\boldsymbol{\xi}_{\text{post}} = \boldsymbol{\Omega}_{\text{post}}\mathbf{r}_0$
$\mathbf{U} = \mathbf{U}_0 = \boldsymbol{\Omega}_{\text{post}}\bar{\mathbf{X}}_0^\intercal$
for t *from* 1 *until convergence* **do**
  for i *from* $\bar{n}_0$ **do**
    ..... Cavity distribution
    $\boldsymbol{\tau}_i = (1 - k_i\bar{\mathbf{x}}_{0[i]}^\intercal\mathbf{u}_i)^{-1}\mathbf{u}_i$
    $\mathbf{r}_{-i} = \mathbf{r}_{\text{EP}} - m_i\bar{\mathbf{x}}_{0[i]}$
    ..... Hybrid distribution
    $s_i = (1 + \bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\tau}_i)^{-1/2}$
    $\gamma_i = s_i\boldsymbol{\tau}_i^\intercal\ \mathbf{r}_{-i}$
    ..... i-th site approximation
    $k_i^{\text{new}} = -\psi_2(\gamma_i)/\big(1 + \bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\tau}_i + \psi_2(\gamma_i)\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\tau}_i\big)$
    $m_i = \psi_1(\gamma_i)s_i + k_i^{\text{new}}\boldsymbol{\tau}_i^\intercal\mathbf{r}_{-i} + k_i^{\text{new}}\psi_1(\gamma_i)s_i\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\tau}_i$
    $\delta k_i = k_i^{\text{new}} - k_i$
    $k_i = k_i^{\text{new}}$
    $\log(1/Z_i) = \log\Phi(\gamma_i) + \frac{1}{2}\log\left(1 + k_i\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\tau}_i\right) + \frac{1}{2}\frac{(\boldsymbol{\tau}_i^\intercal\mathbf{r}_{-i})^2}{\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\tau}_i}$
    $-\frac{1}{2}\frac{\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\tau}_i}{1 + k_i\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\tau}_i}\left(m_i + \frac{\boldsymbol{\tau}_i^\intercal\mathbf{r}_{-i}}{\bar{\mathbf{x}}_{0[i]}^\intercal\boldsymbol{\tau}_i}\right)^2$
    ..... Global approximation
    $\mathbf{r}_{\text{EP}} = \mathbf{r}_{-i} + m_i\bar{\mathbf{x}}_{0[i]}$
    $\mathbf{U} = \mathbf{U} - \mathbf{u}_i\frac{\delta k_i}{1 + \delta k_i\,\bar{\mathbf{x}}_{0[i]}^\intercal\mathbf{u}_i}(\bar{\mathbf{x}}_{0[i]}^\intercal\mathbf{U})$
  end
end
$\boldsymbol{\Omega}_{\text{EP}} = \boldsymbol{\Omega}_{\text{post}} - \mathbf{U}\mathbf{K}\mathbf{U}_0^\intercal$
$\boldsymbol{\xi}_{\text{EP}} = \boldsymbol{\xi}_{\text{post}} + \mathbf{U}_0\mathbf{m} - \mathbf{U}\mathbf{K}\mathbf{U}_0^\intercal\mathbf{r}_{\text{EP}}$
$\log q(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1) = \frac{1}{2}\mathbf{r}^\intercal\boldsymbol{\xi}_{\text{EP}} - \frac{1}{2}\mathbf{r}_0^\intercal\boldsymbol{\xi}_{\text{post}} - \frac{1}{2}\log|\mathbf{I}_{\bar{n}_0} + \mathbf{K}\bar{\mathbf{X}}_0\mathbf{U}_0| + \sum_{i=1}^{\bar{n}_0}\log(1/Z_i)$
Result: ($\boldsymbol{\Omega}_{\text{EP}}, \boldsymbol{\xi}_{\text{EP}}, \log q(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1)$)

$$\boldsymbol{\xi}_{\text{EP}} = \boldsymbol{\Omega}_{\text{EP}}\, \mathbf{r}_{\text{EP}} = \mathbf{Q}_{\text{EP}}^{-1}\mathbf{r}_{\text{EP}} = \boldsymbol{\Omega}\mathbf{r}_{\text{EP}} - \mathbf{U}\mathbf{K}\mathbf{U}_0^\intercal\mathbf{r}_{\text{EP}} = \boldsymbol{\xi}_{\text{post}} + \mathbf{U}_0\mathbf{m} - \mathbf{U}\mathbf{K}\mathbf{U}_0^\intercal\mathbf{r}_{\text{EP}}.$$

We now have all the building blocks to construct the final efficient EP scheme reported below in Algorithm 3

A.1.3 Scalable Gaussian EP for multivariate normal cdf sites

Both the efficient EP schemes from the previous Section can be adapted to the more general case with multivariate normal cdf sites in the target distribution, as it happens in equation (1.26). As before, for ease of notation we replace the superscript "(t_c)" on the updated parameters with "new", and drop the superscript "(t_{c-1})" identifying the outcome of the previous step. Once again, the key for reformulating the EP updates is the application of Woodbury identity, which gives

$$\begin{aligned}
\mathbf{Q}_c^{\text{new}} &= \left(\text{var}_h[\boldsymbol{\beta} \mid \mathbf{y}]\right)^{-1} - \mathbf{Q}_{-c} = \left(\boldsymbol{\Omega}_c + \boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{s}_c^{-1}\boldsymbol{\Pi}_c\mathbf{s}_c^{-1}\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\right)^{-1} - \mathbf{Q}_{-c} \\
&= \boldsymbol{\Omega}_c^{-1} - \bar{\mathbf{X}}_{0[c]}^\intercal(\mathbf{s}_c H_c^{-1}\mathbf{s}_c + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal)^{-1}\bar{\mathbf{X}}_{0[c]} - \boldsymbol{\Omega}_c^{-1} \\
&= \bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{K}_c^{\text{new}}\bar{\mathbf{X}}_{0[c]}
\end{aligned}$$

with $\mathbf{K}_c^{\text{new}} = -(\mathbf{s}_c\boldsymbol{\Pi}_c^{-1}\mathbf{s}_c + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal)^{-1}$, and

$$\begin{aligned}
\mathbf{r}_c^{\text{new}} &= \mathbf{Q}_{\text{EP}}^{\text{new}}\mathbb{E}_h[\boldsymbol{\beta} \mid \mathbf{y}] - \mathbf{r}_{-c} = (\mathbf{Q}_{-c} + \mathbf{Q}_c^{\text{new}})(\boldsymbol{\Omega}_c\mathbf{r}_{-c} + \boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{s}_c^{-1}\boldsymbol{\psi}_c) - \mathbf{r}_{-c} \\
&= \bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{K}_c^{\text{new}}\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\mathbf{r}_{-c} + \bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{K}_c^{\text{new}}\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{s}_c^{-1}\boldsymbol{\psi}_c + \bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{s}_c^{-1}\boldsymbol{\psi}_c \\
&= \bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{K}_c^{\text{new}}\mathbf{s}_c\boldsymbol{\gamma}_c + \bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{K}_c^{\text{new}}\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{s}_c^{-1}\boldsymbol{\psi}_c + \bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{s}_c^{-1}\boldsymbol{\psi}_c \\
&= \bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{m}_c^{\text{new}}
\end{aligned}$$

where $\mathbf{m}_c^{\text{new}} = \left(\mathbf{s}_c^{-1}\boldsymbol{\psi}_c + \mathbf{K}_c^{\text{new}}\mathbf{s}_c\boldsymbol{\gamma}_c + \mathbf{K}_c^{\text{new}}\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{s}_c^{-1}\boldsymbol{\psi}_c\right)$. Accordingly, we can store the $\bar{n}_c$-dimensional vector $\mathbf{m}_c$ and the $\bar{n}_c \times \bar{n}_c$ matrix $\mathbf{K}_c$ for each site $c = 1, \ldots, C$, instead of the full $\mathbf{r}_c$ and $\mathbf{Q}_c$. Accordingly

$$\begin{aligned}
\boldsymbol{\Omega}_c = \mathbf{Q}_{-c}^{-1} &= (\mathbf{Q}_{\text{EP}} - \bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{K}_c\bar{\mathbf{X}}_{0[c]})^{-1} \\
&= \boldsymbol{\Omega}_{\text{EP}} + \boldsymbol{\Omega}_{\text{EP}}\bar{\mathbf{X}}_{0[c]}^\intercal(\mathbf{K}_c^{-1} - \bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_{\text{EP}}\bar{\mathbf{X}}_{0[c]}^\intercal)^{-1}\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_{\text{EP}},
\end{aligned}$$

while

$$\begin{aligned}
\mathbf{Q}_{\text{EP}}^{\text{new}} &= \mathbf{Q}_{-c} + \bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{K}_c^{\text{new}}\bar{\mathbf{X}}_{0[c]} \\
\boldsymbol{\Omega}_{\text{EP}}^{\text{new}} &= \boldsymbol{\Omega}_c + \boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{s}_c^{-1}\boldsymbol{\Pi}_c\mathbf{s}_c^{-1}\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\,.
\end{aligned}$$

For what concerns the update of $\log(1/Z_c)$, instead, recalling that $\boldsymbol{\gamma}_c = \mathbf{s}_c^{-1}\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\mathbf{r}_{-c}$ and introducing $\boldsymbol{\vartheta}_c = (\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal)^{-1}\mathbf{s}_c\boldsymbol{\gamma}_c$ we have

Algorithm 4: EP for Multivariate Normal cdf sites - Quadratic cost per iteration in the number of features $\bar{p}$

$\mathbf{Q}_{\text{EP}} = \mathbf{Q}_0 = \boldsymbol{\Omega}^{-1} + \bar{\mathbf{X}}_1^\intercal \bar{\boldsymbol{\Sigma}}_1^{-1} \bar{\mathbf{X}}_1$
$\mathbf{r}_{\text{EP}} = \mathbf{r}_0 = \boldsymbol{\Omega}^{-1}\boldsymbol{\xi} + \bar{\mathbf{X}}_1^\intercal \bar{\boldsymbol{\Sigma}}_1^{-1} \bar{\mathbf{y}}_1$
for $c = 1, \dots, C$ **do**
  $\mathbf{K}_c = \mathbf{0}_{\bar{n}_c \times \bar{n}_c}$; $\mathbf{m}_c = \mathbf{0}_{\bar{n}_c}$; $\log(1/Z_c) = 0$
end
if $\bar{p} > \bar{n}_1$ **then**
  $\boldsymbol{\Omega}_{\text{EP}} = \boldsymbol{\Omega} - \boldsymbol{\Omega}\bar{\mathbf{X}}_1^\intercal (\bar{\boldsymbol{\Sigma}} + \bar{\mathbf{X}}_1 \boldsymbol{\Omega} X_1^\intercal)^{-1} \bar{\mathbf{X}}_1 \boldsymbol{\Omega}$
else
  $\boldsymbol{\Omega}_{\text{EP}} = (\mathbf{Q}_0)^{-1}$
end
$\boldsymbol{\xi}_{\text{post}} = \boldsymbol{\Omega}_{\text{EP}} \mathbf{r}_0$
for t *from* 1 *until convergence* **do**
  for c *from* C **do**
    Cavity distribution
    $\mathbf{Q}_{-c} = \mathbf{Q}_{\text{EP}} - \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{K}_c \bar{\mathbf{X}}_{0[c]}$
    $\mathbf{r}_{-c} = \mathbf{r} - \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{m}_c$
    Hybrid distribution
    $\boldsymbol{\Omega}_c = \boldsymbol{\Omega}_{\text{EP}} + \boldsymbol{\Omega}_{\text{EP}} \bar{\mathbf{X}}_{0[c]}^\intercal \big(\mathbf{K}_c^{-1} - \bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_{\text{EP}} \bar{\mathbf{X}}_{0[c]}^\intercal\big)^{-1} \bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_{\text{EP}}$
    $\mathbf{s}_c = ([\bar{\boldsymbol{\Sigma}}_{0[c,c]} + \bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal] \odot \mathbf{I}_{\bar{n}_c})^{1/2}$
    $\boldsymbol{\gamma}_c = \mathbf{s}_c^{-1} (\bar{\mathbf{y}}_{0[c]} + \bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c \, \mathbf{r}_{-c})$
    $\boldsymbol{\Gamma}_c = \mathbf{s}_c^{-1} (\bar{\boldsymbol{\Sigma}}_{0[c,c]} + \bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal) \mathbf{s}_c^{-1}$
    Moments evaluation
    $\boldsymbol{\psi}_c = \boldsymbol{\Gamma}_c^{-1} \mathbb{E}[\mathbf{U}_{1c}]$
    $\boldsymbol{\Pi}_c = \boldsymbol{\Gamma}_c^{-1} \big(\text{var}[\mathbf{U}_{1c}] - \boldsymbol{\Gamma}_c\big) \boldsymbol{\Gamma}_c^{-1}$
    c-th site approximation
    $\mathbf{K}_c = -(\mathbf{s}_c \boldsymbol{\Pi}_c^{-1} \mathbf{s}_c + \bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal)^{-1}$
    $\mathbf{m}_c = \mathbf{s}_c^{-1} \boldsymbol{\psi}_c + \mathbf{K}_c \mathbf{s}_c \boldsymbol{\gamma}_c + \mathbf{K}_c \bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{s}_c^{-1} \boldsymbol{\psi}_c$
    $\boldsymbol{\vartheta}_c = (\bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal)^{-1} \mathbf{s}_c \boldsymbol{\gamma}_c$
    $\log(1/Z_c) = \log \Phi_{\bar{n}_c}(\boldsymbol{\gamma}_c, \boldsymbol{\Gamma}_c) + \frac{1}{2} \log |\mathbf{I}_{\bar{n}_c} + \mathbf{K}_c (\bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal)| + \frac{1}{2} \boldsymbol{\gamma}_c^\intercal \mathbf{s}_c \boldsymbol{\vartheta}_c$
      $+ \frac{1}{2} (\mathbf{m}_c + \boldsymbol{\vartheta}_c)^\intercal \mathbf{K}_c^{-1} (\mathbf{s}_c^{-1} \boldsymbol{\Pi}_c \mathbf{s}_c^{-1}) (\bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal) (\mathbf{m}_c + \boldsymbol{\vartheta}_c)$
    Global approximation
    $\mathbf{Q}_{\text{EP}} = \mathbf{Q}_{-c} + \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{K}_c \bar{\mathbf{X}}_{0[c]}$
    $\mathbf{r}_{\text{EP}} = \mathbf{r}_{-c} + \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{m}_c$
    $\boldsymbol{\Omega}_{\text{EP}} = \boldsymbol{\Omega}_c + \boldsymbol{\Omega}_c \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{s}_c^{-1} \boldsymbol{\Pi}_c \mathbf{s}_c^{-1} \bar{\mathbf{X}}_{0[c]} \boldsymbol{\Omega}_c$
  end
end
$\boldsymbol{\xi}_{\text{EP}} = \boldsymbol{\Omega}_{\text{EP}} \, \mathbf{r}$
$\log q(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1) = \frac{1}{2} \mathbf{r}^\intercal \boldsymbol{\xi}_{\text{EP}} - \frac{1}{2} \mathbf{r}_0^\intercal \boldsymbol{\xi}_{\text{post}} - \frac{1}{2} \log |\mathbf{Q}_{\text{EP}}| + \frac{1}{2} \log |\mathbf{Q}_0| + \sum_{c=1}^{C} \log(1/Z_c)$
Result: ($\boldsymbol{\Omega}_{\text{EP}}, \boldsymbol{\xi}_{\text{EP}}, \log q(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1)$)

$$
\begin{aligned}
&= -\frac{1}{2}\mathbf{r}_{\text{EP}}^{\text{new}\intercal}\boldsymbol{\Omega}_{\text{EP}}^{\text{new}}\mathbf{r}_{\text{EP}}^{\text{new}} + \frac{1}{2}\mathbf{r}_{-c}^{\intercal}\boldsymbol{\Omega}_c\mathbf{r}_{-c} \\
&= -\frac{1}{2}(\mathbf{r}_{-c}^{\intercal} + \mathbf{m}_c^{\intercal}\bar{\mathbf{X}}_{0[c]})\left(\boldsymbol{\Omega}_c + \boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal}\mathbf{s}_c^{-1}\boldsymbol{\Pi}_c\mathbf{s}_c^{-1}\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\right)(\mathbf{r}_{-c} + \bar{\mathbf{X}}_{0[c]}^{\intercal}\mathbf{m}_c) + \frac{1}{2}\mathbf{r}_{-c}^{\intercal}\boldsymbol{\Omega}_c\mathbf{r}_{-c} \\
&= -\frac{1}{2}\mathbf{m}_c^{\intercal}(\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal})\left[\mathbf{s}_c^{-1}\boldsymbol{\Pi}_c\mathbf{s}_c^{-1} + (\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal})^{-1}\right](\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal})\mathbf{m}_c \\
&\quad - \mathbf{m}_c^{\intercal}(\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal})\left[\mathbf{s}_c^{-1}\boldsymbol{\Pi}_c\mathbf{s}_c^{-1} + (\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal})^{-1}\right]\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\mathbf{r}_{-c} \\
&\quad - \frac{1}{2}\mathbf{r}_{-c}^{\intercal}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal}\left[\mathbf{s}_c^{-1}\boldsymbol{\Pi}_c\mathbf{s}_c^{-1} + (\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal})^{-1}\right]\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\mathbf{r}_{-c} \\
&\quad + \frac{1}{2}\mathbf{r}_{-c}^{\intercal}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal}(\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal})^{-1}\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\mathbf{r}_{-c} \\
&= \frac{1}{2}(\mathbf{m}_c^{\intercal} + \boldsymbol{\vartheta}_c^{\intercal})(\mathbf{K}_c^{\text{new}})^{-1}(\mathbf{s}_c^{-1}\boldsymbol{\Pi}_c\mathbf{s}_c^{-1})(\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal})(\mathbf{m}_c + \boldsymbol{\vartheta}_c) + \frac{1}{2}(\boldsymbol{\gamma}_c^{\intercal}\mathbf{s}_c\boldsymbol{\vartheta}_c),
\end{aligned}
$$

while

$$
\begin{aligned}
&= \frac{1}{2}\log|\mathbf{Q}_{\text{EP}}^{\text{new}}| - \frac{1}{2}\log|\mathbf{Q}_{-c}| = \frac{1}{2}\log|\mathbf{K}_c^{\text{new}}| + \frac{1}{2}\log|(\mathbf{K}_c^{\text{new}})^{-1} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal}| \\
&= \frac{1}{2}\log|\mathbf{I}_{n_c} + \mathbf{K}_c^{\text{new}}(\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^{\intercal})|
\end{aligned}
$$

Accordingly, we can formulate Algorithm 4 below.

Analogously to the previous Section, we can devise a version of the EP updates with linear cost in the number of covariates even in such multivariate case . Let us define for each block the variables $\mathbf{u}_c = \mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal}$ and $\boldsymbol{\tau}_c = \mathbf{Q}_{-c}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal}$, both with dimension $\bar{p} \times \bar{n}_c$. Accordingly

$$
\begin{aligned}
\boldsymbol{\tau}_c = \mathbf{Q}_{-c}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal} &= (\mathbf{Q}_{\text{EP}} - \bar{\mathbf{X}}_{0[c]}^{\intercal}\mathbf{K}_c\bar{\mathbf{X}}_{0[c]})^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal} \\
&= \mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal} - \mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal}(\mathbf{K}_c^{-1} - \bar{\mathbf{X}}_{0[c]}\mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal})^{-1}\bar{\mathbf{X}}_{0[c]}\mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal} \\
&= \mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal}\left[\mathbf{I}_{\bar{n}_c} + (\mathbf{K}_c^{-1} - \bar{\mathbf{X}}_{0[c]}\mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal})^{-1}\bar{\mathbf{X}}_{0[c]}\mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal}\right] \\
&= \mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal}(\mathbf{K}_c^{-1} - \bar{\mathbf{X}}_{0[c]}\mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal})^{-1}\mathbf{K}_c^{-1} \\
&= \mathbf{u}_c(\mathbf{K}_c^{-1} - \bar{\mathbf{X}}_{0[c]}\mathbf{u}_c)^{-1}\mathbf{K}_c^{-1} = \mathbf{u}_c(\mathbf{I}_{\bar{n}_c} - \mathbf{K}_c\bar{\mathbf{X}}_{0[c]}\mathbf{u}_c)^{-1}.
\end{aligned}
$$

As before, the update of a single site c leads to a change in all $\{\mathbf{u}_s\}_{s=1}^{C}$. Defining the matrix $\delta\mathbf{K}_c = \mathbf{K}_c^{\text{new}} - \mathbf{K}_c$, one has

$$
\begin{aligned}
\mathbf{u}_s^{\text{new}} = \left(\mathbf{Q}_{\text{EP}}^{\text{new}}\right)^{-1}\bar{\mathbf{X}}_{0[s]}^{\intercal} &= (\mathbf{Q}_{\text{EP}} - \mathbf{Q}_c + \mathbf{Q}_c^{\text{new}})^{-1}\bar{\mathbf{X}}_{0[s]}^{\intercal} \\
&= (\mathbf{Q}_{\text{EP}} - \bar{\mathbf{X}}_{0[c]}^{\intercal}\mathbf{K}_c\bar{\mathbf{X}}_{0[c]} + \bar{\mathbf{X}}_{0[c]}^{\intercal}\mathbf{K}_c^{\text{new}}\bar{\mathbf{X}}_{0[c]})^{-1}\bar{\mathbf{X}}_{0[s]}^{\intercal} = (\mathbf{Q}_{\text{EP}} + \bar{\mathbf{X}}_{0[c]}^{\intercal}\delta\mathbf{K}_c\bar{\mathbf{X}}_{0[c]})^{-1}\bar{\mathbf{X}}_{0[s]}^{\intercal} \\
&= \mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[s]}^{\intercal} - \mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal}(\mathbf{I}_{\bar{n}_c} + \delta\mathbf{K}_c\bar{\mathbf{X}}_{0[c]}\mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[c]}^{\intercal})^{-1}\delta\mathbf{K}_c\bar{\mathbf{X}}_{0[c]}\mathbf{Q}_{\text{EP}}^{-1}\bar{\mathbf{X}}_{0[s]}^{\intercal} \\
&= \mathbf{u}_s - \mathbf{u}_c(\mathbf{I}_{\bar{n}_c} + \delta\mathbf{K}_c\bar{\mathbf{X}}_{0[c]}\mathbf{u}_c)^{-1}\delta\mathbf{K}_c\bar{\mathbf{X}}_{0[c]}\mathbf{u}_s.
\end{aligned}
$$

Algorithm 5: Efficient EP for Multivariate Normal cdf sites - Linear cost per iteration in the number of features $\bar{p}$

$\mathbf{Q}_0 = \boldsymbol{\Omega}^{-1} + \bar{\mathbf{X}}_1^\intercal \bar{\boldsymbol{\Sigma}}_1^{-1} \bar{\mathbf{X}}_1$
$\mathbf{r}_0 = \boldsymbol{\Omega}^{-1}\boldsymbol{\xi} + \bar{\mathbf{X}}_1^\intercal \bar{\boldsymbol{\Sigma}}_1^{-1} \bar{\mathbf{y}}_1$
for $c = 1, \dots, C$ **do**
  $\mathbf{K}_c = \mathbf{0}_{\bar{n}_c \times \bar{n}_c}$; $\mathbf{m}_c = \mathbf{0}_{\bar{n}_c}$; $\log(1/Z_c) = 0$
end
if $\bar{p} > \bar{n}_1$ **then**
  $\boldsymbol{\Omega}_{\text{post}} = \boldsymbol{\Omega} - \boldsymbol{\Omega}\bar{\mathbf{X}}_1^\intercal(\bar{\boldsymbol{\Sigma}} + \bar{\mathbf{X}}_1 \boldsymbol{\Omega} X_1^\intercal)^{-1}\bar{\mathbf{X}}_1\boldsymbol{\Omega}$
else
  $\boldsymbol{\Omega}_{\text{post}} = (\mathbf{Q}_0)^{-1}$
end
$\boldsymbol{\xi}_{\text{post}} = \boldsymbol{\Omega}_{\text{post}}\mathbf{r}_0$
$\mathbf{U} = \mathbf{U}_0 = \boldsymbol{\Omega}_{\text{post}}\bar{\mathbf{X}}_0^\intercal$
for t *from* 1 *until convergence* **do**
  for c *from* C **do**
    ..... Cavity distribution
    $\boldsymbol{\tau}_c = \mathbf{u}_c(\mathbf{I}_{\bar{n}_c} - \mathbf{K}_c\bar{\mathbf{X}}_{0[c]}\mathbf{u}_c)^{-1}$
    $\mathbf{r}_{-c} = \mathbf{r} - \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{m}_c$
    ..... Hybrid distribution
    $\mathbf{s}_c = ([\bar{\boldsymbol{\Sigma}}_{0[c,c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal] \odot \mathbf{I}_{\bar{n}_c})^{1/2}$
    $\boldsymbol{\gamma}_c = \mathbf{s}_c^{-1}(\bar{\mathbf{y}}_{0[c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\,\mathbf{r}_{-c})$
    $\boldsymbol{\Gamma}_c = \mathbf{s}_c^{-1}(\bar{\boldsymbol{\Sigma}}_{0[c,c]} + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal)\mathbf{s}_c^{-1}$
    ..... Moments evaluation
    $\boldsymbol{\psi}_c = \boldsymbol{\Gamma}_c^{-1}\mathbb{E}[\mathbf{U}_{1c}]$
    $\boldsymbol{\Pi}_c = \boldsymbol{\Gamma}_c^{-1}\big(\text{var}[\mathbf{U}_{1c}] - \boldsymbol{\Gamma}_c\big)\boldsymbol{\Gamma}_c^{-1}$
    ..... c-th site approximation
    $\mathbf{K}_c^{\text{new}} = -(\mathbf{s}_c\boldsymbol{\Pi}_c^{-1}\mathbf{s}_c + \bar{\mathbf{X}}_{0[c]}\boldsymbol{\tau}_c)^{-1}$
    $\delta\mathbf{K}_c = \mathbf{K}_c^{\text{new}} - \mathbf{K}_c$
    $\mathbf{K}_c = \mathbf{K}_c^{\text{new}}$
    $\mathbf{m}_c = \mathbf{s}_c^{-1}\boldsymbol{\psi}_c + \mathbf{K}_c^{\text{new}}\mathbf{s}_c\boldsymbol{\gamma}_c + \mathbf{K}_c^{\text{new}}\bar{\mathbf{X}}_{0[c]}\boldsymbol{\Omega}_c\bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{s}_c^{-1}\boldsymbol{\psi}_c$
    $\boldsymbol{\vartheta}_c = (\bar{\mathbf{X}}_{0[c]}\boldsymbol{\tau}_c)^{-1}\mathbf{s}_c\boldsymbol{\gamma}_c$
    $\log(1/Z_c) = \log\Phi_{\bar{n}_c}(\boldsymbol{\gamma}_c, \boldsymbol{\Gamma}_c) + \frac{1}{2}\log|\mathbf{I}_{\bar{n}_c} + \mathbf{K}_c(\bar{\mathbf{X}}_{0[c]}\boldsymbol{\tau}_c)| + \frac{1}{2}\boldsymbol{\gamma}_c^\intercal\mathbf{s}_c\boldsymbol{\vartheta}_c$
      $+\frac{1}{2}(\mathbf{m}_c + \boldsymbol{\vartheta}_c)^\intercal\mathbf{K}_c^{-1}(\mathbf{s}_c^{-1}\boldsymbol{\Pi}_c\mathbf{s}_c^{-1})(\bar{\mathbf{X}}_{0[c]}\boldsymbol{\tau}_c)(\mathbf{m}_c + \boldsymbol{\vartheta}_c)$
    ..... Global approximation
    $\mathbf{r}_{\text{EP}} = \mathbf{r}_{-c} + \bar{\mathbf{X}}_{0[c]}^\intercal\mathbf{m}_c$
    $\mathbf{U} = \mathbf{U} + \mathbf{u}_c\big((\mathbf{I}_{\bar{n}_c} + \delta\mathbf{K}_c\bar{\mathbf{X}}_{0[c]}\mathbf{u}_c)^{-1}\delta\mathbf{K}_c\big)(\bar{\mathbf{X}}_{0[c]}\mathbf{U})$
  end
end
$\boldsymbol{\Omega}_{\text{EP}} = \boldsymbol{\Omega}_{\text{post}} - \mathbf{U}\mathbf{K}\mathbf{U}_0^\intercal$
$\boldsymbol{\xi}_{\text{EP}} = \boldsymbol{\xi}_{\text{post}} + \mathbf{U}_0\mathbf{m} - \mathbf{U}\mathbf{K}\mathbf{U}_0^\intercal\mathbf{r}$
$\log q(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1) = \frac{1}{2}\mathbf{r}^\intercal\boldsymbol{\xi}_{\text{EP}} - \frac{1}{2}\mathbf{r}_0^\intercal\boldsymbol{\xi}_{\text{post}} - \frac{1}{2}\log|\mathbf{I}_{\bar{n}_0} + \mathbf{K}\bar{\mathbf{X}}_0\mathbf{U}_0| + \sum_{c=1}^{C}\log(1/Z_c)$
Result: ($\boldsymbol{\Omega}_{\text{EP}}, \boldsymbol{\xi}_{\text{EP}}, \log q(\bar{\mathbf{y}}_0 \mid \bar{\mathbf{y}}_1)$)

Grouping $\{\mathbf{u}_c\}_{c=1}^C$ into a single $\bar{p} \times \bar{n}_0$ matrix $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_C)$, the cost of EP updates scales again linearly in the number of features $\bar{p}$ since

$$\mathbf{U}^{\text{new}} = \mathbf{U} + \mathbf{u}_c \left[(\mathbf{I}_{\bar{n}_c} + \delta \mathbf{K}_c \bar{\mathbf{X}}_{0[c]} \mathbf{u}_c)^{-1} \delta \mathbf{K}_c \right] (\bar{\mathbf{X}}_{0[c]} \mathbf{U}) \ .$$

Eventually, the final matrix inversion an the determinant calculation can be performed as seen before in the case of univariate truncated-normal sites. Indeed, we still have that $\mathbf{Q}_{\text{EP}} = \mathbf{Q}_0 + \sum_{c=1}^C \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{K}_c \bar{\mathbf{X}}_{0[c]} = \mathbf{Q}_0 + \bar{\mathbf{X}}_0^\intercal \mathbf{K} \bar{\mathbf{X}}_0$ and $\mathbf{r} = \mathbf{r}_0 + \sum_{c=1}^C \bar{\mathbf{X}}_{0[c]}^\intercal \mathbf{m}_c = \bar{\mathbf{X}}_0^\intercal \mathbf{M}$, where now $\mathbf{K} = \text{diag}(\{\mathbf{K}_c\}_{c=1}^C)$ is a $\bar{n}_0 \times \bar{n}_0$ block-diagonal matrix, while $\mathbf{m} = (\mathbf{m}_1^\intercal, \dots, \mathbf{m}_C^\intercal)^\intercal$ is a vector of length $\bar{n}_0$. Accordingly, as before

$$\begin{aligned} \boldsymbol{\Omega}_{\text{EP}} &= \mathbf{Q}_{\text{EP}}^{-1} = \boldsymbol{\Omega}_0 - \mathbf{U}\mathbf{K}\mathbf{U}_0^\intercal \\ \boldsymbol{\xi}_{\text{EP}} &= \boldsymbol{\xi}_{\text{post}} + \mathbf{Q}_{\text{EP}}^{-1} \mathbf{r}_{\text{EP}} = \mathbf{U}_0 \mathbf{m} - \mathbf{U}\mathbf{K}\mathbf{U}_0^\intercal \mathbf{r}, \end{aligned}$$

and

$$-\frac{1}{2} \log |\mathbf{Q}_{\text{EP}}| + \frac{1}{2} \log |\mathbf{Q}_0| = -\frac{1}{2} \log |\mathbf{I}_{\bar{n}_0} + \mathbf{K}\bar{\mathbf{X}}_0 \mathbf{U}_0|.$$

The resulting efficient EP schemes in summarized in Algorithm 5.

A.2 PLQ optimality and semismooth coordinate descent

A.2.1 Optimality of the PLQ bound

In this Section we prove equation (2.20) from Chapter 2, which states the optimality of the PLQ bound with the family $\mathcal{H}_{\text{S}}(\zeta)$, given that $\zeta \neq 0$. With no loss of generality and for ease of exposition, we hereby consider a translated version $\tilde{h}(r)$ of the target function $h(r)$ from equation (2.8)

$$\tilde{h}(r) = h(r) - h(0) = -\log \cosh(r/2) \ ,$$

so that $\tilde{h}(0) = 0$, while we still have $\tilde{h}(-r) = \tilde{h}(r)$. Accordingly, every element $h_{\text{S}}(r \mid \zeta) \in \mathcal{H}_{\text{S}}(\zeta)$ gives a proper minorizer for the adjusted target by simply applying the same rigid translation $\tilde{h}_{\text{S}}(r \mid \zeta) = h_{\text{S}}(r \mid \zeta) - h(0)$, so that

$$\begin{cases} \tilde{h}(r) \geq \tilde{h}_{\text{S}}(r \mid \zeta) = \tilde{h}(\zeta) + a_{\text{S}}(\zeta)(r - \zeta) - \frac{1}{2} w_{\text{S}}(\zeta)(r^2 - \zeta^2) - \nu_{\text{S}}(\zeta)(|r| - |\zeta|) \\ \tilde{h}(\zeta) = \tilde{h}_{\text{S}}(\zeta \mid \zeta) \ . \end{cases}$$

In particular, the minorization requirement implies

$$\begin{cases} \tilde{h}(\zeta) = \tilde{h}(-\zeta) \geq \tilde{h}_{\text{S}}(-\zeta \mid \zeta) = \tilde{h}(\zeta) - 2a_{\text{S}}(\zeta)\zeta \\ 0 = \tilde{h}(0) \geq \tilde{h}_{\text{S}}(0 \mid \zeta) = \tilde{h}(\zeta) - a_{\text{S}}(\zeta)\zeta + \frac{1}{2} w_{\text{S}}(\zeta)\zeta^2 + \nu_{\text{S}}(\zeta)|\zeta| \ . \end{cases}$$

Additionally, since we have assumed that $\zeta \neq 0$, the minorizers will be differentiable at ζ. As such, the tangent minorization requirement further gives

$$\tilde{h}'(\zeta) = \tilde{h}'_{\text{S}}(\zeta \mid \zeta) = a_{\text{S}}(\zeta) - w_{\text{S}}(\zeta)\zeta - \nu_{\text{S}}(\zeta)\,\text{sign}(\zeta)$$

where $\tilde{h}'(r) = \partial\tilde{h}'(r)/\partial r$. Combining this with the previous conditions, we have that

$$\begin{cases} a_{\text{S}}(\zeta) - \nu_{\text{S}}(\zeta)\,\text{sign}(\zeta) = \tilde{h}'(\zeta) + w_{\text{S}}(\zeta)\zeta \\ w_{\text{S}}(\zeta) \geq \frac{2}{\zeta^2}\Big(\tilde{h}(\zeta) - \tilde{h}'(\zeta)\,\zeta\Big) \\ a_{\text{S}}(\zeta)\zeta \geq 0\,. \end{cases}$$

Conversely, we recall that the PLQ bound $\tilde{h}_{\text{PLQ}}(r \mid \zeta)$ arises by imposing both that $a_{\text{PLQ}}(\zeta) = 0$, which implies symmetry with respect to the origin, and that $\tilde{h}(0) = \tilde{h}_{\text{PLQ}}(0 \mid \zeta)$. This translates into

$$\begin{aligned} w_{\text{PLQ}}(\zeta) &= \frac{2}{\zeta^2}\Big(\tilde{h}(\zeta) - \tilde{h}'(\zeta)\,\zeta\Big) \\ \nu_{\text{PLQ}}(\zeta) &= \frac{1}{|\zeta|}\Big(\tilde{h}'(\zeta)\,\zeta - 2\,\tilde{h}(\zeta)\Big)\,, \end{aligned}$$

which in particular means that $w_{\text{PLQ}}(\zeta) \leq w_{\text{S}}(\zeta)$. Assume now that $\zeta > 0$. If $r > 0$, then

$$\begin{aligned} \tilde{h}_{\text{PLQ}}(r \mid \zeta) - \tilde{h}_{\text{S}}(r \mid \zeta) &= -\Big(\nu_{\text{PLQ}}(\zeta) - \nu_{\text{S}}(\zeta) + a_{\text{S}}(\zeta)\Big)(r - \zeta) - \frac{1}{2}\Big(w_{\text{PLQ}}(\zeta) - w_{\text{S}}(\zeta)\Big)(r^2 - \zeta^2) \\ &= \Big(w_{\text{PLQ}}(\zeta) - w_{\text{S}}(\zeta)\Big)\zeta(r - \zeta) - \frac{1}{2}\Big(w_{\text{PLQ}}(\zeta) - w_{\text{S}}(\zeta)\Big)(r^2 - \zeta^2) \\ &= -\frac{1}{2}\Big(w_{\text{PLQ}}(\zeta) - w_{\text{S}}(\zeta)\Big)(r - \zeta)^2 \geq 0\,. \end{aligned}$$

Conversely, if $r < 0$

$$\tilde{h}_{\text{PLQ}}(r \mid \zeta) - \tilde{h}_{\text{S}}(r \mid \zeta) = \Big(\tilde{h}_{\text{PLQ}}(-r \mid \zeta) - \tilde{h}_{\text{S}}(-r \mid \zeta)\Big) - 2\,a_{\text{S}}(\zeta)\,r \geq 0\,.$$

Indeed, the first term is non-negative thanks to the previous equation, while the second is one is non-negative because $r < 0$ and $a_{\text{S}}(\zeta) \geq 0$, since $\zeta > 0$. As by definition $\tilde{h}_{\text{PLQ}}(0 \mid \zeta) = \tilde{h}(0) \geq \tilde{h}_{\text{S}}(0 \mid \zeta)$, we have that $\tilde{h}_{\text{PLQ}}(r \mid \zeta) \geq \tilde{h}_{\text{S}}(r \mid \zeta)$ for any $r \in \Re$, while analogous results can be derived for the case $\zeta < 0$.

A.2.2 Semi-smooth coordinate-wise updates

In this Section, we report the coordinate-wise updates for the semi-smooth surrogate from equation (2.23), which follow directly from an extension of the construction by Yi & Huang (2017). A key aspect of the methodology by the authors is that it updates simultaneously

a regression coefficient β_j and its corresponding subgradient $v_j \in \partial|\beta_j|$ at each step of the coordinate-wise optimization. Indeed, the solutions to the univariate optimization problems can be tackled by addressing the corresponding Karush-Kuhn-Tucker conditions (Rockafellar, 1970), thanks to the concavity of the objective function. Given the current values for the primal and dual $s^{(t,s)} = (\boldsymbol{\beta}^{(t,s)}, \mathbf{v}^{(t,s)})^\intercal$ variables, the updates for intercept β_0, usually not affected by the penalization, take the form

$$\beta_0^{(t,s+1)} \leftarrow \tilde{\beta}_0^{(t)} - \frac{\sum_{i=1}^n \big(-(y_i - 1/2) + w_{\text{PLQ},i}^{(t)} \cdot \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t,s)} + \nu_{\text{PLQ},i}^{(t)} \cdot d'_\epsilon(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t,s)}) \big)}{\sum_{i=1}^n \big(w_{\text{PLQ},i}^{(t)} + \nu_{\text{PLQ},i}^{(t)} \cdot \mathbb{1}(|\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t,s)}| \leq \epsilon)/\epsilon \big)},$$

where $\mathbb{1}(|r| \leq \epsilon)/\epsilon$ provides a valid subgradient of $d'_\epsilon(r) \coloneqq \partial d_\epsilon(r)/\partial r$ (Yi & Huang, 2017). Conversely, the joint updates for $s_j^{(t,s+1)} = (\beta_j^{(t,s+1)}, v_j^{(t,s+1)})^\intercal$ takes a two-fold form, depending on the value of $|\beta_j^{(t,s)} + v_j^{(t,s)}|$. In particular, if $|\beta_j^{(t,s)} + v_j^{(t,s)}| > 1$

$$\begin{aligned}
\beta_j^{(t,s+1)} \leftarrow \beta_j^{(t,s)} &- \frac{\sum_{i=1}^n \big(-(y_i - 1/2) + w_{\text{PLQ},i}^{(t)} \cdot \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t,s)} + \nu_{\text{PLQ},i}^{(t)} \cdot d'_\epsilon(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t,s)}) \big) x_{ij}}{\sum_{i=1}^n \big(w_{\text{PLQ},i}^{(t)} + \nu_{\text{PLQ},i}^{(t)} \cdot \mathbb{1}(|\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t,s)}| \leq \epsilon)/\epsilon \big) x_{ij}^2 + \lambda(1-\alpha)} \\
&+ \frac{\lambda(1-\alpha)\beta_j^{(t,s)} + \lambda\, \alpha\, \text{sign}(\beta_j^{(t,s)} + v_j^{(t,s)})}{\sum_{i=1}^n \big(w_{\text{PLQ},i}^{(t)} + \nu_{\text{PLQ},i}^{(t)} \cdot \mathbb{1}(|\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t,s)}| \leq \epsilon)/\epsilon \big) x_{ij}^2 + \lambda(1-\alpha)} \\
v_j^{(t,s+1)} \leftarrow \text{sign}(\beta_j^{(t,s)} &+ v_j^{(t,s)}) \,,
\end{aligned}$$

while if $|\tilde{\beta}_j + \tilde{v}_j| \leq 1$

$$\begin{aligned}
\beta_j^{(t,s+1)} &\leftarrow 0 \\
v_j^{(t,s+1)} &\leftarrow -\frac{\sum_{i=1}^n \big(-(y_i - 1/2) + w_{\text{PLQ},i}^{(t)} \cdot \mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t,s)} + \nu_{\text{PLQ},i}^{(t)} \cdot d'_\epsilon(\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t,s)}) \big) x_{ij}}{\lambda\alpha} \\
&\quad + \frac{\beta_j^{(t,s)} \sum_{i=1}^n \big(w_{\text{PLQ},i}^{(t)} + \nu_{\text{PLQ},i}^{(t)} \cdot \mathbb{1}(|\mathbf{x}_i^\intercal \boldsymbol{\beta}^{(t,s)}| \leq \epsilon)/\epsilon \big) x_{ij}^2}{\lambda\alpha} \,.
\end{aligned}$$

A.3 PLQ-VB updates and efficient computations

A.3.1 Variational parameters updates for the PLQ bound

In the current Section we prove equation (3.11) from Chapter 3, which gives the update rule for the variational parameters $\{\zeta_i\}_{i=1}^n$ under the piece-wise quadratic approximation introduced in Section 3.3. The goal is to maximize the function

$$\begin{aligned}
Q_{\text{PLQ}}(\boldsymbol{\zeta} \mid \boldsymbol{\zeta}^{(t)}) &= \mathbb{E}_{\bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})} \big[\log \bar{p}_{\text{PLQ}}(\boldsymbol{\beta}, \mathbf{y} \mid \boldsymbol{\zeta}) \big] \\
&= \mathbb{E}_{\bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})} \Big[\log p(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^n \log \bar{p}_{\text{PLQ}}(y_i \mid \boldsymbol{\beta}, \zeta_i) \Big]
\end{aligned}$$

over $\boldsymbol{\zeta} \in \Re^n$, where for ease of notation we are going to write $q^{(t+1)}(\boldsymbol{\beta}) = \bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}^{(t)})$. Notice that $Q_{\text{PLQ}}(\boldsymbol{\zeta} \mid \boldsymbol{\zeta}^{(t)})$ depends on the variational parameters only via their absolute value $\{|\zeta_i|\}_{i=1}^n$, so that we can restrict our attention to the maximization over $\varrho_i = |\zeta_i|$, for all $i = 1, \dots, n$. Accordingly, $Q_{\text{PLQ}}(\boldsymbol{\varrho} \mid \boldsymbol{\varrho}^{(t)}) = \sum_{i=1}^n Q_i(\varrho_i \mid \boldsymbol{\varrho}^{(t)})$, with

$$\begin{aligned} Q_i(\varrho_i \mid \boldsymbol{\varrho}^{(t)}) = &- \log\cosh(\varrho_i/2) - \nu_{\text{PLQ}}(\varrho_i)\Big(\mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})} \left[|\mathbf{x}_i^\intercal \boldsymbol{\beta}| \right] - \varrho_i \Big) \\ &- \frac{1}{2} w_{\text{PLQ}}(\varrho_i)\Big(\mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})} \left[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 \right] - \varrho_i^2 \Big) + \text{const} \end{aligned}$$

and we want to solve $Q_i'(\varrho_i \mid \boldsymbol{\varrho}^{(t)}) \coloneqq \partial Q_i(\varrho_i \mid \boldsymbol{\varrho}^{(t)}) / \partial \varrho_i = 0$ for all $i = 1, \dots, n$. Recall that

$$\begin{aligned} \nu_{\text{PLQ}}(\varrho_i) &= \frac{2}{\varrho_i} \left(\log\cosh(\varrho_i/2) - \frac{1}{4} \varrho_i \tanh(\varrho_i/2) \right) \\ w_{\text{PLQ}}(\varrho_i) &= \frac{2}{\varrho_i^2} \left(- \log\cosh(\varrho_i/2) + \frac{1}{2} \varrho_i \tanh(\varrho_i/2) \right), \end{aligned}$$

and $\nu_{\text{PLQ}}(\varrho_i) + \varrho_i w_{\text{PLQ}}(\varrho_i) = \frac{1}{2} \tanh(\varrho_i/2)$. Simple algebraic calculations allow to obtain

$$\begin{aligned} \nu_{\text{PLQ}}'(\varrho_i) &\coloneqq \frac{\partial \nu_{\text{PLQ}}(\varrho_i)}{\partial \varrho_i} = w_{\text{PLQ}}(\varrho_i) - \frac{1}{4} \operatorname{sech}^2(\varrho_i/2) \\ w_{\text{PLQ}}'(\varrho_i) &\coloneqq \frac{\partial w_{\text{PLQ}}(\varrho_i)}{\partial \varrho_i} = -\frac{2}{\varrho_i} \nu_{\text{PLQ}}'(\varrho_i). \end{aligned}$$

Accordingly

$$\begin{aligned} Q_i'(\varrho_i \mid \boldsymbol{\varrho}^{(t)}) = &-\frac{1}{2} \tanh(\varrho_i/2) + \nu_{\text{PLQ}}(\varrho_i) + \varrho_i w_{\text{PLQ}}(\varrho_i) - \nu_{\text{PLQ}}'(\varrho_i)\Big(\mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})} \left[|\mathbf{x}_i^\intercal \boldsymbol{\beta}| \right] - \varrho_i \Big) \\ &- \frac{1}{2} w_{\text{PLQ}}'(\varrho_i)\Big(\mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})} \left[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 \right] - \varrho_i^2 \Big). \end{aligned}$$

The first three terms cancel out as a consequence of equation (2.19), so that

$$\begin{aligned} Q_i'(\varrho_i \mid \boldsymbol{\varrho}^{(t)}) &= -\nu_{\text{PLQ}}'(\varrho_i) \, \mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})} \left[|\mathbf{x}_i^\intercal \boldsymbol{\beta}| \right] - \frac{1}{2} w_{\text{PLQ}}'(\varrho_i) \, \mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})} \left[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 \right] \\ &\quad + \frac{1}{2} \varrho_i^2 \Big(w_{\text{PLQ}}'(\varrho_i) + \frac{2}{\varrho_i} \nu_{\text{PLQ}}'(\varrho_i) \Big) \\ &= \frac{1}{2} w_{\text{PLQ}}'(\varrho_i) \Big(\varrho_i \cdot \mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})} \left[|\mathbf{x}_i^\intercal \boldsymbol{\beta}| \right] - \mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})} \left[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 \right] \Big), \end{aligned}$$

which is clearly equal to zero for $\varrho_i = \mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})} \left[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 \right] / \mathbb{E}_{q^{(t+1)}(\boldsymbol{\beta})} \left[|\mathbf{x}_i^\intercal \boldsymbol{\beta}| \right]$. In turn, this gives

$$\left| \zeta_i^{(t+1)} \right| = \frac{\mathbb{E}_{\bar{p}_{\text{PLQ}}(\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\zeta}^{(t)})} \left[(\mathbf{x}_i^\intercal \boldsymbol{\beta})^2 \right]}{\mathbb{E}_{\bar{p}_{\text{PLQ}}(\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\zeta}^{(t)})} \left[|\mathbf{x}_i^\intercal \boldsymbol{\beta}| \right]} \qquad i = 1, \dots, n$$

as in equation (3.11).

A.3.2 Approximate posterior moments and scalable implementation

In the current Section we provide the details for the scalable implementation of the PLQ-VB posterior, under the scale-mixture representation for the Laplace contributions, as in equation (3.13)

$$\bar{p}_{\text{PLQ}}(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\zeta}) = \frac{1}{2^n} \int_{(\Re^+)^n} \frac{\phi_p\big(\boldsymbol{\xi}_0; \boldsymbol{\Omega}_0\big)}{\phi_p\big(\boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa}); \boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa})\big)} \prod_{i=1}^{n} p(\kappa_i)\, d\kappa_i = \frac{1}{2^n}\, \mathbb{E}_{p(\boldsymbol{\kappa})} \left[\, \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \,\right] \tag{17}$$

We begin by noticing that the weights $\varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa})$ can be rewritten as

$$\begin{aligned} \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa}) &= \phi_p\big(\boldsymbol{\xi}_0; \boldsymbol{\Omega}_0\big) / \phi_p\big(\boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa}); \boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa})\big) \\ &= \frac{|\boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa})|^{1/2}}{|\boldsymbol{\Omega}_0|^{1/2}} \frac{\exp\left\{\frac{1}{2}\boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa})^\intercal \boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa})\right\}}{\exp\left\{\frac{1}{2}\boldsymbol{\xi}_0^\intercal \boldsymbol{\Omega}_0 \boldsymbol{\xi}_0\right\}} \end{aligned}$$

where Woodbury Identity gives us

$$\begin{aligned} \frac{|\boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa})|^{1/2}}{|\boldsymbol{\Omega}_0|^{1/2}} = \left| \text{diag}\left(\left\{ \frac{\kappa_i}{w_{\text{PLQ}}(\zeta_i)\kappa_i + \nu^2_{\text{PLQ}}(\zeta_i)} \right\}_{i=1}^{n} \right) + \mathbf{X}\boldsymbol{\Omega}_0\mathbf{X}^\intercal \right|^{-1/2} \cdot \\ \cdot \prod_{i=1}^{n} \left(\frac{\kappa_i}{w_{\text{PLQ}}(\zeta_i)\kappa_i + \nu^2_{\text{PLQ}}(\zeta_i)} \right)^{1/2}, \end{aligned}$$

and

$$\begin{aligned} &= \boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa})^\intercal \boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \\ &= \mathbf{r}^\intercal \boldsymbol{\Omega}_0 \mathbf{r} - \mathbf{r}^\intercal \boldsymbol{\Omega}_0 \mathbf{X}^\intercal \left(\text{diag}\left(\left\{ \frac{\kappa_i}{w_{\text{PLQ}}(\zeta_i)\kappa_i + \nu^2_{\text{PLQ}}(\zeta_i)} \right\}_{i=1}^{n} \right) + \mathbf{X}\boldsymbol{\Omega}_0\mathbf{X}^\intercal \right)^{-1} \mathbf{X}\boldsymbol{\Omega}_0 \mathbf{r} \quad , \end{aligned}$$

where we have introduced $\mathbf{r} = (\boldsymbol{\Omega}_0^{-1}\boldsymbol{\xi}_0 + \mathbf{X}^\intercal(\mathbf{y} - 1/2\,\mathbf{1}_n))$. For what concerns the approximate posterior mean of $\boldsymbol{\beta}$, it is easy to very that

$$\mathbb{E}_{p_{\text{PLQ}}(\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\zeta})} \left[\, \boldsymbol{\beta} \,\right] = \frac{\mathbb{E}_{p(\boldsymbol{\kappa})} \left[\, \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa})\, \boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \,\right]}{\mathbb{E}_{p(\boldsymbol{\kappa})} \left[\, \varpi(\boldsymbol{\beta}, \boldsymbol{\kappa}) \,\right]} \,,$$

while, in accordance with the law of total variance $\text{var}[\boldsymbol{\beta}] = \mathbb{E}[\text{var}[\boldsymbol{\beta} \mid \boldsymbol{\kappa}]] + \text{var}[\mathbb{E}[\boldsymbol{\beta} \mid \boldsymbol{\kappa}]]$, for the approximate covariance of $\boldsymbol{\beta}$ we have

$$\begin{aligned} \text{var}_{p_{\text{PLQ}}(\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\zeta})} \left[\, \boldsymbol{\beta} \,\right] = &\frac{\mathbb{E}_{p(\boldsymbol{\kappa})} \left[\, \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa})\, \boldsymbol{\Omega}(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \,\right]}{\mathbb{E}_{p(\boldsymbol{\kappa})} \left[\, \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \,\right]} + \frac{\mathbb{E}_{p(\boldsymbol{\kappa})} \left[\, \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa})\, \boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa})\, \boldsymbol{\xi}^\intercal(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \,\right]}{\mathbb{E}_{p(\boldsymbol{\kappa})} \left[\, \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \,\right]} \\ &- \frac{\mathbb{E}_{p(\boldsymbol{\kappa})} \left[\, \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa})\, \boldsymbol{\xi}(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \,\right]}{\mathbb{E}_{p(\boldsymbol{\kappa})} \left[\, \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \,\right]} \cdot \frac{\mathbb{E}_{p(\boldsymbol{\kappa})} \left[\, \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa})\, \boldsymbol{\xi}^\intercal(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \,\right]}{\mathbb{E}_{p(\boldsymbol{\kappa})} \left[\, \varpi(\boldsymbol{\zeta}, \boldsymbol{\kappa}) \,\right]} \,. \end{aligned}$$

Bibliography

AGRESTI, A. (2013). *Categorical Data Analysis (Third Edition)*. Wiley.

ALBERT, J. & CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

ALBERT, J. & CHIB, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics* **57**, 829–836.

ALI, A. & TIBSHIRANI, R. J. (2019). The generalized lasso problem and uniqueness. *Electronic Journal of Statistics* **13**, 2307 – 2347.

AMEMIYA, T. (1984). Tobit models: A survey. *Journal of Econometrics* **24**, 3–61.

ANDRIEU, C. & DOUCET, A. (2002). Particle filtering for partially observed Gaussian state space models. *Journal of the Royal Statistical Society: Series B* **64**, 827–836.

ARELLANO-VALLE, R. B. & AZZALINI, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* **33**, 561–574.

ARELLANO-VALLE, R. B. & AZZALINI, A. (2021). Some properties of the unified skew-normal distribution. *Statistical Papers* , 1–27.

ARELLANO-VALLE, R. B., BRANCO, M. D. & GENTON, M. G. (2006). A unified view on skewed distributions arising from selections. *Canadian Journal of Statistics* **34**, 581–601.

ARNOLD, B. C. & BEAVER, R. J. (2000). Hidden truncation models. *Sankhyā: The Indian Journal of Statistics, Series A* **62**, 23–35.

ARNOLD, B. C., BEAVER, R. J., AZZALINI, A., BALAKRISHNAN, N., BHAUMIK, A., DEY, D., CUADRAS, C. & SARABIA, J. M. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Test* **11**, 7–54.

ARNOLD, T. B. & TIBSHIRANI, R. J. (2016). Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics* **25**, 1–27.

AZZALINI, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171–178.

AZZALINI, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics* **32**, 159–188.

AZZALINI, A. & BACCHIERI, A. (2010). A prospective combination of phase II and phase III in drug development. *Metron* **68**, 347–369.

AZZALINI, A. & CAPITANIO, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B* **61**, 579–602.

AZZALINI, A. & CAPITANIO, A. (2013). *The Skew-Normal and Related Families*. Cambridge University Press.

AZZALINI, A. & DALLA VALLE, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–726.

BAZÁN, J. L., BOLFARINE, H. & BRANCO, M. D. (2010). A framework for skew-probit links in binary regression. *Communications in Statistics-Theory and Methods* **39**, 678–697.

BENAVOLI, A., AZZIMONTI, D. & PIGA, D. (2020). Skew Gaussian processes for classification. *Machine Learning* **109**, 1877–1902.

BENAVOLI, A., AZZIMONTI, D. & PIGA, D. (2021). A unified framework for closed–form nonparametric regression, classification, preference and mixed problems with skew Gaussian processes. *Machine Learning* **110**, 3095–3133.

BHADRA, A., DATTA, J., POLSON, N. G. & WILLARD, B. T. (2020). Global-local mixtures: A unifying framework. *Sankhya A* **82**, 426–447.

BISHOP, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

BISHOP, C. M. & SVENSÉN, M. (2003). Bayesian hierarchical mixtures of experts. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*.

BLEI, D. M., KUCUKELBIR, A. & MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877.

BLISS, C. I. (1934). The method of probits. *Science* **79**, 38–39.

BÖHNING, D. (1992). Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics* **44**, 197–200.

BÖHNING, D. & LINDSAY, B. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics* **40**, 641–663.

BOTEV, Z. I. (2017). The normal law under linear restrictions: Simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B* **88**, 125–148.

BROWNE, R. P. & MCNICHOLAS, P. D. (2015). Multivariate sharp quadratic bounds via $\boldsymbol{\Sigma}$-strong convexity and the Fenchel connection. *Electronic Journal of Statistics* **9**, 1913 – 1938.

CAMPBELL, T. & LI, X. (2019). Universal boosting variational inference. In *Advances in Neural Information Processing Systems*, vol. 32.

CANALE, A., PAGUI, E. C. K. & SCARPA, B. (2016). Bayesian modeling of university first-year students' grades after placement test. *Journal of Applied Statistics* **43**, 3015–3029.

CANDÈS, E. & SUR, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics* **48**, 27 – 42.

CAO, J., DURANTE, D. & GENTON, M. G. (2022). Scalable computation of predictive probabilities in probit models with Gaussian process priors. *Journal of Computational and Graphical Statistics* , 1–12.

CAO, J., GENTON, M. G., KEYES, D. E. & TURKIYYAH, G. M. (2019). Hierarchical-block conditioning approximations for high-dimensional multivariate normal probabilities. *Statistics and Computing* **29**, 585–598.

CAO, J., GENTON, M. G., KEYES, D. E. & TURKIYYAH, G. M. (2021). Exploiting low-rank covariance structures for computing high-dimensional normal and student-t probabilities. *Statistics and Computing* **31**, 1–16.

CARBONETTO, P. & STEPHENS, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* **7**, 73–108.

CARVALHO, C. M., POLSON, N. G. & SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

CHAN, J. C.-C. & JELIAZKOV, I. (2009). MCMC estimation of restricted covariance matrices. *Journal of Computational and Graphical Statistics* **18**, 457–480.

CHEN, M.-H., DEY, D. K. & SHAO, Q.-M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association* **94**, 1172–1186.

CHIB, S. (1992). Bayes inference in the tobit censored regression model. *Journal of Econometrics* **51**, 79 – 99.

CHIB, S. & GREENBERG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.

CHIB, S., GREENBERG, E. & JELIAZKOV, I. (2009). Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics* **18**, 321–348.

CHIB, S. & JELIAZKOV, I. (2006). Inference in semiparametric dynamic models for binary longitudinal data. *Journal of the American Statistical Association* **101**, 685–700.

CHIPMAN, H. A., GEORGE, E. I. & MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **4**, 266–298.

CHOPIN, N. (2011). Fast simulation of truncated Gaussian distributions. *Statistics and Computing* **21**, 275–288.

CHOPIN, N. & PAPASPILIOPOULOS, O. (2020). *An introduction to sequential Monte Carlo*. Springer.

CHOPIN, N. & RIDGWAY, J. (2017). Leave pima indians alone: Binary regression as a benchmark for Bayesian computation. *Statistical Science* **32**, 64–87.

CONSONNI, G. & MARIN, J.-M. (2007). Mean-field variational approximate Bayesian inference for latent variable models. *Computational Statistics & Data Analysis* **52**, 790–798.

CRAIG, P. (2008). A new reconstruction of multivariate normal orthant probabilities. *Journal of the Royal Statistical Society: Series B* **70**, 227–243.

CUI, Y., CHANG, T.-H., HONG, M. & PANG, J.-S. (2020). A study of piecewise linear-quadratic programs. *J. Optim. Theory Appl.* **186**, 523–553.

DE LEEUW, J. (1977). Applications of convex analysis to multidimensional scaling. In *Recent Developments in Statistics*. North Holland Publishing Company, pp. 133–146.

DE LEEUW, J. & HEISER, W. J. (1977). Convergence of correction matrix algorithms for multidimensional scaling. *Geometric Representations of Relational data* **36**, 735–752.

DE LEEUW, J. & LANGE, K. (2009). Sharp quadratic majorization in one dimension. *Computational Statistics & Data Analysis* **53**, 2471–2484.

DE OLIVEIRA, V. (2005). Bayesian inference and prediction of Gaussian random fields based on censored data. *Journal of Computational and Graphical Statistics* **14**, 95–115.

DEHAENE, G. & BARTHELMÉ, S. (2015). Bounding errors of expectation-propagation. In *Advances in Neural Information Processing Systems*, vol. 28.

DEHAENE, G. & BARTHELMÉ, S. (2018). Expectation propagation in the large data limit. *Journal of the Royal Statistical Society: Series B* **80**, 199–217.

DEMARIS, A. (2004). *Regression with Social Data: Modeling Continuous and Limited Response Variables*. John Wiley & Sons.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B* **39**, 1–22.

DUA, D. & GRAFF, C. (2017). UCI machine learning repository. http://archive.ics.uci.edu/ml.

DURANTE, D. (2019). Conjugate bayes for probit regression via unified skew-normal distributions. *Biometrika* **106**, 765–779.

DURANTE, D. & RIGON, T. (2019). Conditionally conjugate mean-field variational Bayes for logistic models. *Statistical Science* **34**, 472 – 485.

ERMIS, B. & BOUCHARD, G. (2014). Iterative splits of quadratic bounds for scalable binary tensor factorization. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*.

FASANO, A. & DURANTE, D. (2022). A class of conjugate priors for multinomial probit models which includes the multivariate normal one. *Journal of Machine Learning Research* **23**, 1–26.

FASANO, A., DURANTE, D. & ZANELLA, G. (2022). Scalable and accurate variational Bayes for high-dimensional binary regression models. *Biometrika* **In press**.

FASANO, A. & REBAUDO, G. (2021). Variational inference for the smoothing distribution in dynamic probit models. In *Book of Short Papers of the Italian Statistical Society*.

FASANO, A., REBAUDO, G., DURANTE, D. & PETRONE, S. (2021). A closed-form filter for binary time series. *Statistics and Computing* **31**, 1–20.

FRIEDMAN, J., HASTIE, T., HOFLING, H. & TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**, 302 – 332.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33 1**, 1–22.

FRÜHWIRTH-SCHNATTER, S. & FRÜHWIRTH, R. (2007). Auxiliary mixture sampling with applications to logistic models. *Computational Statistics & Data Analysis* **51**, 3509–3528.

FRÜHWIRTH-SCHNATTER, S., FRÜHWIRTH, R., HELD, L. & RUE, H. (2009). Improved auxiliary mixture sampling for hierarchical models of non-gaussian data. *Statistics and Computing* **19**, 479–492.

GALARZA MORALES, C. E., MATOS, L. A., DEY, D. K. & LACHOS, V. H. (2021). On moments of folded and doubly truncated multivariate extended skew-normal distributions. *Journal of Computational and Graphical Statistics* , 1–29.

GASSMANN, H. (2003). Multivariate normal probabilities: Implementing an old idea of Plackett's. *Journal of Computational and Graphical Statistics* **12**, 731–752.

GELFAND, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association* **95**, 1300–1304.

GELMAN, A., JAKULIN, A., G., P. M. & SU, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* **2**, 1360–1383.

GENTON, M., KEYES, D. & TURKIYYAH, G. (2018). Hierarchical decompositions for the computation of high-dimensional multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **27**, 268–277.

GENZ, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **1**, 141–149.

GENZ, A. (2004). Numerical computation of rectangular bivariate and trivariate normal and *t* probabilities. *Statistics and Computing* **14**, 251–260.

GENZ, A. & BRETZ, F. (2002). Comparison of methods for the computation of multivariate *t* probabilities. *Journal of Computational and Graphical Statistics* **11**, 950–971.

GENZ, A. & BRETZ, F. (2009). *Computation of Multivariate Normal and t Probabilities*, vol. 195. Springer.

GESSNER, A., KANJILAL, O. & HENNIG, P. (2020). Integrals over Gaussians under linear domain constraints. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, vol. 108.

GHOSH, I., BHATTACHARYA, A. & PATI, D. (2022). Statistical optimality and stability of tangent transform algorithms in logit models. *Journal of Machine Learning Research* **23**.

GIORDANO, R. J., BRODERICK, T. & JORDAN, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational bayes. In *Advances in Neural Information Processing Systems*, vol. 28.

GIROLAMI, M. & ROGERS, S. (2006). Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation* **18**, 1790–1817.

GONZÁLEZ-FARIAS, G., DOMINGUEZ-MOLINA, A. & GUPTA, A. K. (2004). Additive properties of skew normal random vectors. *Journal of Statistical Planning and Inference* **126**, 521–534.

GRAMACY, R. B. & POLSON, N. G. (2012). Simulation-based regularized logistic regression. *Bayesian Analysis* **7**, 567 – 590.

GREENE, W. (2008). *Econometric Analysis*. Pearson Education India.

GUNAWARDANA, A. & BYRNE, W. (2005). Convergence theorems for generalized alternating minimization procedures. *Journal of Machine Learning Research* **6**, 2049–2073.

GUO, F., WANG, X., FAN, K., BRODERICK, T. & DUNSON, D. B. (2016). Boosting variational inference. *arXiv preprint arXiv:1611.05559* .

GUPTA, A. K., AZIZ, M. A. & NING, W. (2013). On some properties of the unified skew normal distribution. *Journal of Statistical Theory and Practice* **7**, 480–495.

GUPTA, A. K., GONZÁLEZ-FARIAS, G. & DOMINGUEZ-MOLINA, J. A. (2004). A multivariate skew normal distribution. *Journal of Multivariate Analysis* **89**, 181–190.

HAARIO, H., SAKSMAN, E. & TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242.

HANS, C. (2009). Bayesian lasso regression. *Biometrika* **96**, 835–845.

HASTIE, T., MONTANARI, A., ROSSET, S. & TIBSHIRANI, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics* **50**, 949 – 986.

HASTIE, T., TIBSHIRANI, R. & WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

HAUSMAN, J. & WISE, D. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* **46**, 403–426.

HAYTER, A. & LIN, Y. (2013). The evaluation of trivariate normal probabilities defined by linear inequalities. *Journal of Statistical Computation and Simulation* **83**, 668–676.

HOERL, A. E. & KENNARD, R. W. (1981). Ridge regression - 1980: Advances, algorithms, and applications. *American Journal of Mathematical and Management Sciences* **1**, 5–83.

HOFFMAN, M. & BLEI, D. (2015). Stochastic structured variational inference. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, vol. 38.

HOFFMAN, M. D., BLEI, D. M., WANG, C. & PAISLEY, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research* **14**, 1303–1347.

HOFFMAN, M. D. & GELMAN, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623.

HOLMES, C. & MALLICK, B. (2001). Bayesian regression with multivariate linear splines. *Journal of the Royal Statistical Society: Series B* **63**, 3–17.

HOLMES, C. C. & HELD, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145–168.

HORRACE, W. (2005). Some results on the multivariate truncated normal distribution. *Journal of Multivariate Analysis* **94**, 209–221.

HUNTER, D. R. & LANGE, K. (2004). A tutorial on mm algorithms. *The American Statistician* **58**, 30–37.

HUTTON, J. & STANGHELLINI, E. (2011). Modelling bounded health scores with censored skew-normal distributions. *Statistics in Medicine* **30**, 368–376.

IMAI, K. & VAN DYK, D. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics* **124**, 311–334.

JAAKKOLA, T. & JORDAN, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**, 25–37.

JAMES, J. (2017). MM algorithm for general mixed multinomial logit models. *Journal of Applied Econometrics* **32**, 841–857.

JELIAZKOV, I., GRAVES, J. & KUTZBACH, M. (2008). Fitting and comparison of models for multivariate ordinal outcomes. In *Bayesian Econometrics*, vol. 23. Emerald Group Publishing Limited, pp. 115–156.

JOHNDROW, J., DUNSON, D. & LUM, K. (2013). Diagonal orthant multinomial probit models. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, vol. 31.

JOHNDROW, J. E., SMITH, A., PILLAI, N. & DUNSON, D. B. (2018). MCMC for imbalanced categorical data. *Journal of the American Statistical Association* **114**, 1394–1403.

JONES, B., CARVALHO, C., DOBRA, A., HANS, C., CARTER, C. & WEST, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science* **20**, 388–400.

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

KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal Basic Engineering* **82**, 35–45.

KHAN, M. E. E., BOUCHARD, G., MURPHY, K. P. & MARLIN, B. M. (2010). Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Processing Systems*, vol. 23.

KIM, S.-J., KOH, K., BOYD, S. & GORINEVSKY, D. (2009). ℓ_1 trend filtering. *SIAM Review* **51**, 339–360.

KING, B. & KOWAL, D. R. (2021). Warped dynamic linear models for time series of counts. *arXiv preprint arXiv:2110.14790* .

KLOPFENSTEIN, Q., BERTRAND, Q., GRAMFORT, A., SALMON, J. & VAITER, S. (2020). Model identification and local linear convergence of coordinate descent. *arXiv preprint arXiv:2010.11825* .

KOWAL, D. R. (2021). Conjugate priors for count and rounded data regression. *arXiv preprint arXiv:2110.12316* .

KOZUMI, H. & KOBAYASHI, G. (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation* **81**, 1565–1578.

KRISHNAPURAM, B., CARIN, L., FIGUEIREDO, M. & HARTEMINK, A. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, 957–968.

KUCUKELBIR, A., TRAN, D., RANGANATH, R., GELMAN, A. & BLEI, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research* **18**, 1–45.

KULLBACK, S. & LEIBLER, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79 – 86.

KUSS, M., RASMUSSEN, C. E. & HERBRICH, R. (2005). Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research* **6**, 1679–1704.

LANG, S. & BREZGER, A. (2004). Bayesian p-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.

LANGE, K. (2016). *MM Optimization Algorithms*. SIAM.

LANGE, K., WON, J.-H., LANDEROS, A. & ZHOU, H. (2021). *Nonconvex Optimization via MM Algorithms: Convergence Theory*. John Wiley & Sons, Ltd.

LEE, S., HUANG, J. & HU, J. (2010). Sparse logistic principal components analysis for binary data. *The Annals of Applied Statistics* **4 3**, 1579–1601.

LI, Q., LIN, N. & XI, R. (2010). Bayesian regularized quantile regression. *Bayesian Analysis* **5**, 533 – 556.

LI, Y. & ARCE, G. R. (2004). A maximum likelihood approach to least absolute deviation regression. *EURASIP J. Adv. Signal Process* **2004**, 1762–1769.

LI, Y. & TURNER, R. E. (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, vol. 29.

LOAIZA-MAYA, R., SMITH, M. S., NOTT, D. J. & DANAHER, P. J. (2021). Fast and accurate variational inference for models with many latent variables. *Journal of Econometrics* , In press.

LUCET, Y., BAUSCHKE, H. H. & TRIENIS, M. (2009). The piecewise linear-quadratic model for computational convex analysis. *Computational Optimization and Applications* **43**, 95–118.

MA, J., XU, L. & JORDAN, M. I. (2000). Asymptotic convergence rate of the em algorithm for Gaussian mixtures. *Neural Computation* **12**, 2881–2907.

MANRIQUE, A. & SHEPHARD, N. (1998). Simulation-based likelihood inference for limited dependent processes. *The Econometrics Journal* **1**, 174–202.

MARLIN, B., KHAN, M. & MURPHY, K. (2011). Piecewise bounds for estimating Bernoulli-logistic latent Gaussian models. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*.

MCCULLOCH, R., POLSON, N. & ROSSI, P. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* **99**, 173–193.

MCCULLOCH, R. & ROSSI, P. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* **64**, 207–240.

MCLACHLAN, G. & KRISHNAN, T. (1996). *The EM algorithm and extensions*. Wiley.

MENG, X.-L. (1994). On the rate of convergence of the ECM algorithm. *The Annals of Statistics* **22**, 326–339.

MENG, X.-L. & RUBIN, D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267–278.

MILLER, A. C., FOTI, N. J. & ADAMS, R. P. (2017). Variational boosting: Iteratively refining posterior approximations. In *Proceedings of the 34th International Conference on Machine Learning*, vol. 70.

MINKA, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, vol. 17.

MIWA, T., HAYTER, A. & KURIKI, S. (2003). The evaluation of general non-centred orthant probabilities. *Journal of the Royal Statistical Society: Series B* **65**, 223–234.

NAVEAU, P., GENTON, M. G. & SHEN, X. (2005). A skewed kalman filter. *Journal of Multivariate Analysis* **94**, 382–400.

NEAL, R. M. & HINTON, G. E. (1998). *A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants*. Springer, pp. 355–368.

NESTEROV, Y. & NEMIROVSKII, A. (1994). *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics.

NICKISCH, H. & RASMUSSEN, C. E. (2008). Approximations for binary Gaussian process classification. *Journal of Machine Learning Research* **9**, 2035–2078.

NISHIMURA, A., ZHANG, Z. & SUCHARD, M. (2021). Hamiltonian zigzag sampler got more momentum than its markovian counterpart: Equivalence of two zigzags under a momentum refreshment limit. *arXiv preprint arXiv:2104.07694* .

NOBILE, A. (1998). A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statistics and Computing* **8**, 229–242.

NOMURA, N. (2016). Evaluation of Gaussian orthant probabilities based on orthogonal projections to subspaces. *Statistics and Computing* **26**, 187–197.

OHISHI, M., FUKUI, K., OKAMURA, K., ITOH, Y. & YANAGIHARA, H. (2021). Coordinate optimization for generalized fused lasso. *Communications in Statistics - Theory and Methods* **50**, 5955–5973.

ORMEROD, J. & WAND, M. (2010). Explaining variational approximations. *The American Statistician* **64**, 140 – 153.

PAKMAN, A. & PANINSKI, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics* **23**, 518–542.

PARIZI, S. N., HE, K., AGHAJANI, R., SCLAROFF, S. & FELZENSZWALB, P. (2019). Generalized majorization-minimization. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97.

PARK, T. & CASELLA, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.

PARK, T. & VAN DYK, D. A. (2009). Partially collapsed Gibbs samplers: Illustrations and applications. *Journal of Computational and Graphical Statistics* **18**, 283–305.

POLSON, N. G. & SCOTT, J. G. (2013). Data augmentation for non-Gaussian regression models using variance-mean mixtures. *Biometrika* **100**, 459–471.

POLSON, N. G., SCOTT, J. G. & WINDLE, J. (2012). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association* **108**, 1339 – 1349.

QIN, Q. & HOBERT, J. P. (2019). Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression. *The Annals of Statistics* **47**, 2320–2347.

RANGANATH, R., GERRISH, S. & BLEI, D. M. (2014). Black box variational inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, vol. 33.

RASMUSSEN, C. E. & WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

REN, L., DU, L., CARIN, L. & DUNSON, D. (2011). Logistic stick-breaking process. *Journal of Machine Learning Research* **12 Jan**, 203–239.

RIDGWAY, J. (2016). Computation of Gaussian orthant probabilities in high dimension. *Statistics and Computing* **26**, 899–916.

RIIHIMÄKI, J., JYLÄNKI, P. & VEHTARI, A. (2014). Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *Journal of Machine Learning Research* **14**, 75–109.

ROBERTS, G. O. & ROSENTHAL, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* **16**, 351–367.

ROCKAFELLAR, R. T. (1970). *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press.

RODRÍGUEZ, A. & DUNSON, D. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6**, 145–178.

RUE, H., MARTINO, S. & CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B* **71**, 319–392.

SAHA, A. & TEWARI, A. (2013). On the nonasymptotic convergence of cyclic coordinate descent methods. *SIAM Journal on Optimization* **23**, 576–601.

SAHU, S. K., DEY, D. K. & BRANCO, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics* **31**, 129–150.

SCOTT, J. G. & SUN, L. (2013). Expectation-maximization for logistic regression. *arXiv preprint arXiv:1306.0040* .

SOYER, R. & SUNG, M. (2013). Bayesian dynamic probit models for the analysis of longitudinal data. *Computational Statistics & Data Analysis* **68**, 388–398.

STERN, S. (1992). A method for smoothing simulated moments of discrete probabilities in multinomial probit models. *Econometrica* **60**, 943–952.

SUR, P. & CANDÈS, E. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences* **116**, 14516–14525.

TALHOUK, A., DOUCET, A. & MURPHY, K. (2012). Efficient Bayesian inference for multivariate probit models with sparse inverse correlation matrices. *Journal of Computational and Graphical Statistics* **21**, 739–757.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58**, 267–288.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. & KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B* **67**, 91–108.

TIBSHIRANI, R. J. & TAYLOR, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics* **39**, 1335–1371.

TOBIN, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26**, 24–36.

TRINH, G. & GENZ, A. (2015). Bivariate conditioning approximations for multivariate normal probabilities. *Statistics and Computing* **25**, 989–996.

TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications* **109**, 475–494.

TUTZ, G. (1991). Sequential models in categorical regression. *Computational Statistics & Data Analysis* **11**, 275–295.

VEHTARI, A., GELMAN, A., SIVULA, T., JYLÄNKI, P., TRAN, D., SAHAI, S., BLOMSTEDT, P., CUNNINGHAM, J., SCHIMINOVICH, D. & ROBERT, C. (2020). Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data. *Journal of Machine Learning Research* **21**, 1–53.

VORONIN, S., OZKAYA, G. & YOSHIDA, D. (2015). Convolution based smooth approximations to the absolute value function with application to non-smooth regularization. *arXiv preprint arXiv:1408.6795* .

WAINWRIGHT, M., JAAKKOLA, T. & WILLSKY, A. (2005). A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory* **51**, 2313–2335.

WAINWRIGHT, M. J. & JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1**, 1–305.

WANG, B. & TITTERINGTON, D. M. (2004). Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*.

WU, T. T. & LANGE, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics* **2**, 224 – 244.

WU, T. T. & LANGE, K. (2010). The MM alternative to EM. *Statistical Science* **25**, 492 – 505.

YEDIDIA, J. S., FREEMAN, W. & WEISS, Y. (2000). Generalized belief propagation. In *Advances in Neural Information Processing Systems*, vol. 13.

YI, C. & HUANG, J. (2017). Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics* **26**, 547–557.

YU, D., WON, J.-H., LEE, T., LIM, J. & YOON, S. (2015). High-dimensional fused lasso regression using majorization–minimization and parallel processing. *Journal of Computational and Graphical Statistics* **24**, 121–153.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques* .

ZENS, G., FRÜHWIRTH-SCHNATTER, S. & WAGNER, H. (2020). Ultimate pólya–gamma samplers - efficient MCMC for possibly imbalanced binary and categorical data. *arXiv preprint arXiv:2011.06898* .

ZHANG, H., ZHOU, P., YANG, Y. & FENG, J. (2019). Generalized majorization-minimization for non-convex optimization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*.

ZHANG, Z., ARELLANO-VALLE, R. B., GENTON, M. G. & HUSER, R. (2021a). Tractable Bayes of skew-elliptical link models for correlated binary data. *arXiv preprint arXiv:2101.02233* .

ZHANG, Z., NISHIMURA, A., BASTIDE, P., JI, X., PAYNE, R. P., GOULDER, P., LEMEY, P. & SUCHARD, M. A. (2021b). Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models. *The Annals of Applied Statistics* **15**, 230–251.

ZHAO, T., LIU, H. & ZHANG, T. (2018). Pathwise coordinate optimization for sparse learning: Algorithm and theory. *The Annals of Statistics* **46**, 180 – 218.

ZHOU, H. & ZHANG, Y. (2012). EM vs MM: A case study. *Computational Statistics & Data Analysis* **56**, 3909–3920.

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67**, 301–320.