UNIVERSITÀ COMMERCIALE “LUIGI BOCCONI”

PhD SCHOOL

PhD program in Statistics

Cycle: XXXII

Disciplinary Field: SECS-S/01

Advances in Bayesian Inference for Binary and Categorical Data

Advisor: Prof. Daniele Durante

Co-Advisor: Prof. Igor Prünster

PhD Thesis by

Augusto Fasano

ID number: 3029881

Year 2021

Abstract

Bayesian binary probit regression and its extensions to time-dependent observations and multi-class responses are popular tools in binary and categorical data regression due to their high interpretability and non-restrictive assumptions. Although the theory is well established in the frequentist literature, such models still face a florid research in the Bayesian framework. This is mostly due to the fact that state-of-the-art methods for Bayesian inference in such settings are either computationally impractical or inaccurate in high dimensions and in many cases a closed-form expression for the posterior distribution of the model parameters is, apparently, lacking. The development of improved computational methods and theoretical results to perform inference with this vast class of models is then of utmost importance.

In order to overcome the above-mentioned computational issues, we develop a novel variational approximation for the posterior of the coefficients in high-dimensional probit regression with binary responses and Gaussian priors, resulting in a unified skew-normal (SUN) approximating distribution that converges to the exact posterior as the number of predictors p increases. Moreover, we show that closed-form expressions are actually available for posterior distributions arising from models that account for correlated binary time-series and multi-class responses. In the former case, we prove that the filtering, predictive and smoothing distributions in dynamic probit models with Gaussian state variables are, in fact, available and belong to a class of SUN distributions whose parameters can be updated recursively in time via analytical expressions, allowing to develop an i.i.d. sampler together with an optimal sequential Monte Carlo procedure. As for the latter case, i.e. multi-class probit models, we show that many different formulations developed in the literature in separate ways admit a unified view and a closed-form SUN posterior distribution under a SUN prior distribution (thus including the Gaussian case). This allows to implement computational methods which outperform state-of-the-art routines in high-dimensional settings by leveraging SUN properties and the variational methods introduced for the binary probit.

Finally, motivated also by the possible linkage of some of the above-mentioned models to the Bayesian nonparametrics literature, a novel species-sampling model for partially-exchangeable observations is introduced, with the double goal of both predicting the class (or species) of the future observations and testing for homogeneity among the different available populations. Such model arises from a combination of Pitman-Yor processes and leverages on the appealing features of both hierarchical and nested structures developed in the Bayesian nonparametrics literature. Posterior inference is feasible thanks to the implementation of a marginal Gibbs sampler, whose pseudo-code is given in full detail.

Acknowledgements

The completion of present thesis would have certainly not been possible without the great mentorship of my two thesis advisors, Prof. Daniele Durante and Prof. Igor Prünster. I would then like to express my most sincere gratitude to Daniele for constantly supporting my Ph.D. research with contagious enthusiasm, which allowed me to truly enjoy doing research, in addition to continuously motivating me to do my best. I knew I could count on him and on his countless ideas when I needed a helping hand to figure out in which direction I should move for my research. On the other hand, I would certainly not be here if it was not for Igor, to whom I am deeply grateful for motivating me to pursue a Ph.D. and for guiding me across research since I was an Allievo at Collegio Carlo Alberto in Turin: his patience, his immense knowledge and his guidance across years were fundamental to get to the end of this Ph.D. thesis, which I hope will be a new beginning. I am certain I could have not hoped to have better Ph.D. supervisors.

I should also thank Giacomo Zanella for his great availability and for all the things he made me learn each time I passed by his office. Moreover, I am also extremely grateful to my Ph.D. colleague and friend Giovanni Rebaudo: thanks to him I could have a lot of fun in doing research, and I hope we will enjoy research together also in the future. Last but not least, I could have hardly pursued this Ph.D. without the support of my parents, to whom it goes my sincere gratitude.

Contents

Chapter 1

Introduction

Regression models for dichotomous and categorical data are facing a considerable interest, due to their wide range of possible applications, spanning across many fields of research (Agresti, 2013, 2018). Such models are suited to study how the probability mass function of a categorical response variable y—that is, a variable whose measurement scale consists in a set of categories—changes with a set of observed predictors $\mathbf{x} \in \mathbb{R}^p$. Such categorical outcomes are widespread in health sciencies, social and political sciences, econometrics, machine learning and statistics in general. Just to mention a few examples of practical relevance (see Agresti (2018) for additional applications), y could represent the severity of an injury ("none", "mild", "moderate", "severe"), or the type of a tumor mass, including both the dichotomous case "benignant" vs "malignant" or more granular outcomes as "benignant", "malignant type 1", "malignant type 2". In social sciences applications, where the interest is in measuring attitudes and opinions, categorical responses regression models are used to study political orientation ("Democrat" vs "Republican" or multi-class) as well as preferences among different alternatives that customers face, in the so-called discrete-choice scenarios (see Greene (2003) for a detailed overview). The range of other possible applications is huge, as they can be used in standard classification tasks, like email spam detection ("spam" vs "legitimate mail"), handwritten digit classification (Rasmussen and Williams, 2006), or to model the occurrence of a certain event (yes, no), like the presence of a certain disease or the rise of a stock price in a trading day.

Although the frequentist theory is now well-established (Agresti, 2013), Bayesian inference for binary and categorical data regression models is still an open area of research, both from a methodological and computational standpoint. Many of these models, including the ones considered in the present thesis, admit an interpretation in terms of latent (unobserved) continuous data, which can be used to get further insights on theoretical properties of the model at hand and to perform Bayesian computations (Albert and Chib, 1993, 2001; Consonni and Marin, 2007; Andrieu and Doucet, 2002; McCulloch and Rossi, 1994; Imai and Van Dyk, 2005; Girolami and Rogers, 2006; Girolami and Zhong, 2007).

In particular, the present thesis focuses on computational and methodological advances in Bayesian inference for the binary probit model, together with more sophisticated constructions to account for correlated binary time series and categorical data. Considering for the moment the probit model for binary outcomes for ease of exposition, each observation y_i is a Bernoulli random variable with mean parameter $\text{pr}(y_i = 1 \mid \boldsymbol{\beta}) = \Phi(\mathbf{x}_i^\intercal \boldsymbol{\beta})$, being $\boldsymbol{\beta} \in \mathbb{R}^p$ the parameter vector of the effects of each covariate on the output. Such a model admits a dual interpretation in terms of partially observed latent Gaussian random variables: it can indeed be rewritten as $y_i = \mathbb{1}[z_i > 0]$, with $z_i \sim \text{N}(\mathbf{x}_i^\intercal \boldsymbol{\beta}, 1)$ and $\mathbb{1}[\cdot]$ the indicator function. See Chapter 2 for further details.

Such latent variables can be used as auxiliary variables or can have an interpretable meaning: in the voting example reported above, they can for instance represent a continuous measure of voter i's preference towards one of the two parties, ending up in her/him voting that party if such continuous-valued measure falls above zero. Similar concepts of underlying continuous latent variables driving the observations are present also when one considers extensions of the probit model to account for dependent binary time series or multi-class responses. Such models are studied in detail in Chapters 3 and 4 of the present thesis, respectively: the corresponding latent continuous-valued process is a multivariate dynamic linear model (Petris et al., 2009) in the former case and a set of unobserved utilities, one for each possible choice, in the latter case (Greene, 2003).

As apparent from the previous arguments, all these models arise from a hierarchical construction. It comes then with no surprise that Bayesian hierarchical models have been widely used in such contexts, as they represent a natural tool to interpret the model construction and perform posterior inference. However, the apparent lack of a conjugate prior distribution for all these classes of probit models motivated a rich literature for computational methods in order to perform Bayesian inference in these settings, due to their central role in binary and categorical Bayesian data analysis. See Chapters 2, 3 and 4 and references therein for more detailed literature reviews. Most of the available methods rely on Markov-Chain Monte Carlo (MCMC) methods to sample from the posterior distribution, exploiting the above-mentioned hierarchical constructions to develop a Gibbs sampler (Albert and Chib, 1993; Holmes and Held, 2006; Pakman and Paninski, 2014; Imai and Van Dyk, 2005). Approximate methods have also been developed for the binary and categorical probit models (Consonni and Marin, 2007; Chopin and Ridgway, 2017; Girolami and Rogers, 2006), while state-of-the-art sequential Monte-Carlo (SMC) routines (Andrieu and Doucet, 2002) provide the standard tool for online inference in the univariate binary time-series setting. However, the available MCMC methods are impractical in large p scenarios, and approximate methods either suffer the same computational problems (Chopin and Ridgway, 2017) or are inaccurate (Consonni and Marin, 2007).

A first solution to these methodological and computational bottlenecks was recently given by Durante (2019), who showed a conjugacy result for the probit model for binary

data with Gaussian priors on the coefficients. This finding, in addition to provide a deeper theoretical understanding of the posterior distribution of the model parameters, allows to implement computational strategies for posterior inference that outperform stat-of-the-art routines in the large p small n scenario, though leaving the computation or approximation of the posterior an open research question when p is large and n is moderate-to-large. Such a problem is tackled in Chapter 2, where a novel variational approximation of the posterior distribution of the binary probit model under Gaussian priors is developed. Durante (2019) showed that, within such a framework, the posterior distribution for the p probit coefficients has a unified skew-normal (SUN) kernel (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014), which can be expressed via a convolution of a p-variate normal and an n-variate truncated normal with full covariance matrix. As the latter part is the main reason for the computational inefficiency as n increases, we propose a variational approximation for the SUN posterior distribution, which factorizes the multivariate truncated normal density via a product of univariate truncated normal densities. Such a result can be formally interpreted as a partially factorized mean-field variational Bayes strategy (Bishop, 2006; Blei et al., 2017) which provides a tighter approximation to the posterior distribution for the probit coefficients, compared to state-of-the-art solutions in Bayesian variational inference (Consonni and Marin, 2007), while crucially preserving skewness. Such a method is proven to be asymptotically exact as the number of covariates p diverges: in such a case, the Kullback-Leibler divergence (Kullback and Leibler, 1951) between the variational approximation and the exact posterior goes to zero with probability one.

Motivated by the above-mentioned methodological and computational advances, other interesting research questions involve the extension of such results to the frameworks of time-dependent binary observations and multi-class probit models mentioned above. We address these questions in Chapters 3 and 4, respectively. Considering the binary time series case, studied in Chapter 3, it admits an interpretation in terms of a partially-observed dynamic linear model (Petris et al., 2009; Durbin and Koopman, 2012). In this framework, when one considers usual Gaussian-Gaussian dynamic linear models, all the distributions of interest are available in closed-form, thanks to the well-known Kalman filter (Kalman, 1960). However, when we move to the non-Gaussian binary case, such a routine is lacking in the literature, leaving it an open research question, which we successfully tackle. Indeed, even though SMC methods are efficient in performing online inference in binary time-series analysis, they are still sub-optimal with respect to closed-form expressions or exact sampling methods. Moreover, they suffer from the well-known problem of particle degeneracy (Durbin and Koopman, 2012) when one moves from online, i.e. filtering and predictive, to batch, i.e. smoothing, inference, leaving the joint and marginal smoothing an open research question. In Chapter 3, we prove that the filtering, predictive and smoothing distributions of dynamic probit models with Gaussian

state variables belong to the class of unified skew-normals (SUN) and that a closed-form expression for the observation predictive probability is available. Leveraging on SUN properties (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014; Durante, 2019), we propose methods to draw independent and identically distributed (i.i.d.) samples from the joint smoothing distribution, which can easily be adapted to obtain i.i.d. samples from filtering and predictive distributions, thereby improving state-of-the-art approximate or sequential Monte Carlo inference in small-to-moderate dimensional dynamic models. A scalable and optimal (in the sense of Doucet et al. (2000)) particle filter which exploits SUN properties is also developed in order to deal with online inference in large dimensional dynamic models.

Extensions of the binary probit model do not stop to binary time series, but include a large class of models for multi-class outputs. Many different constructions, with associated MCMC procedures, have been proposed in the literature (Hausman and Wise, 1978; Tutz, 1991; Stern, 1992; McCulloch and Rossi, 1994; McCulloch et al., 2000; Albert and Chib, 2001; Imai and Van Dyk, 2005). However, a unified view on them is lacking at the moment, as well as a closed-form expression for the corresponding posterior distributions. The goal of Chapter 4 is indeed to provide such a unified view, together with theoretical and computational advances, on models for categorical data that can be formulated as extensions of the binary probit model, generally referred to as Multinomial Probit (MNP) models. We focus in particular on the original formulation by Hausman and Wise (1978), on an alternative model with class-specific parameters (Stern, 1992) and on a sequential construction arising from initial formulations by Tutz (1991) and Albert and Chib (2001). We show that all the three models, originally developed in a separate way for different kinds of data, lead to a SUN posterior distribution of the parameters under a SUN (and hence also Gaussian) prior distribution, developing an efficient sampling procedure which outperforms state-of-the-art methods in the large p moderate n scenario. Such results are then used as a starting point to develop variational inference techniques extending the routine introduced in Chapter 2, allowing to get posterior estimates when both p and n are large, with particular focus on the case $p > n$.

All the models discussed so far deal with probabilistic classification, so that predictions about future observations will be given in the form of class probabilities and not only as point class predictions: class guesses can then be obtained as solutions of a decision problem, after the specification of a loss-function. A research area strictly related to this probabilistic classification framework is given by the Bayesian nonparametrics approach to species sampling problems, introduced in Lijoi et al. (2007), where the main interest is in prediction of additional observations, conditionally on the available data. In particular, key quantities to predict are the number of new species in an additional sample, which can be seen as a measure of species diversity, or the rate of decay of the probability of discovering new species. Frequentist analogs originated in connection with ecological

problems (Good, 1953; Good and Toulmin, 1956). Since then, however, they have been applied in various other fields, so that the term 'species' has actually gained a metaphoric meaning and can indicate different possible types, genes, agents or categories, depending on the context. In particular, these models are facing an increasing interest in the Bayesian nonparametrics community, with a wide range of applications spreading across genetics (Lijoi et al., 2007; Favaro et al., 2009, 2012), economics (Lijoi et al., 2016) and machine learning (Teh, 2006). See also De Blasi et al. (2015) for an extensive overview. In this framework, the most used model is arguably the Pitman Yor Process (PYP) (Pitman and Yor, 1997), being it preferred to the popular Dirichlet Process (DP) (Ferguson, 1973) mainly due to the fact that the probability that a new observation forms a new cluster, conditionally on the available sample, depends on the number of already created clusters, providing greater flexibility than the DP, where such probability depends only on the overall sample size. This is also reflected in different asymptotic distributions for the number of observed clusters as the population size diverges, with the PYP showing a power-law behaviour, which is common in many empirical studies (Mitzenmacher, 2004; Goldwater et al., 2006), contrary to the logarithmic growth observed for the DP, which appears too restrictive.

When moving to the partially exchangeable framework, where multiple samples are collected across different, but related, studies, Bayesian hierarchical models have proved to be an effective tool, since they naturally allow to borrow information across groups (Teh et al., 2006; Teh, 2006; Teh and Jordan, 2010; Camerlenghi et al., 2019b). Such hierarchical constructions, although quite flexible, do not allow to have ties of distributions of various groups, so that these will have different, but related, distributions. A popular practice to specify a model that allows for ties among distributions of different groups is to exploit nested structures (Rodríguez et al., 2008; Camerlenghi et al., 2019a). However, these models either suffer from a degeneracy issue that does not allow ties in the observations across different groups without degenerating to the exchangeable case (Rodríguez et al., 2008), or are computationally infeasible for more than two groups (Camerlenghi et al., 2019a). For these reasons, none of them is suitable to perform population homogeneity testing in species sampling models with more than two groups.

Motivated by this methodological and computational lack, in Chapter 5 we introduce a novel species-sampling model for the multiple-sample setting, allowing predictive inference of future observations as well as clustering of the different population distributions, so to perform population homogeneity testing. Such model arises by combining PYPs in a way to exploit the advantages of both the hierarchical and nested constructions developed in the Bayesian nonparametrics literature (Teh et al., 2006; Teh, 2006; Rodríguez et al., 2008; Teh and Jordan, 2010; Camerlenghi et al., 2019a,b).

In order to do so, we extend the Hierarchical Pitman-Yor process (HPYP) (Teh, 2006) by adding a latent structure on the population distributions, so that we allow ties across

them and we can perform both the above-mentioned tasks at the same time. We then show that the distribution of the induced random partition admits a closed-form expression, which allows to gain a deeper insight on the theoretical properties of the model. Posterior inference is feasible thanks to an MCMC routine which allows to evaluate the functionals of interest and to perform homogeneity testing for different populations with multiple groups.

Chapter 2

Variational Bayes for High-Dimensional Probit Models

2.1 Introduction

The absence of tractable posterior distributions in several Bayesian models, and the recent abundance of high-dimensional datasets have motivated a growing interest in strategies for scalable learning of approximate posteriors, beyond classical sampling-based Markov chain Monte Carlo (MCMC) methods (e.g., Green et al., 2015). Deterministic approximations, such as variational Bayes (VB) (Blei et al., 2017) and expectation-propagation (EP) (Minka, 2001), provide powerful approaches to improve computational efficiency in posterior inference. However, in high-dimensional models these methods still face open problems in terms of scalability and quality of the posterior approximation. Notably, such issues also arise in basic predictor-dependent models for binary responses (Agresti, 2013), which are routinely used and provide a building block in several hierarchical models (e.g., Chipman et al., 2010; Rodriguez and Dunson, 2011). Recalling a recent review by Chopin and Ridgway (2017), the problem of posterior computation in binary regression is particularly challenging when the number of predictors p becomes large. Indeed, while standard sampling-based algorithms and deterministic approximations can easily deal with small p problems, these strategies are impractical when p is large; e.g., $p > 1000$.

Classical specifications of Bayesian regression models for binary data assume that the dichotomous responses $y_i \in \{0; 1\}$, $i = 1, \ldots, n$, are conditionally independent realizations from a Bernoulli variable $\text{Bern}[g(\mathbf{x}_i^\intercal \boldsymbol{\beta})]$, given a fixed p-dimensional vector of predictors $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal \in \mathbb{R}^p$, $i = 1, \ldots, n$, and the associated coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal \in \mathbb{R}^p$. The mapping $g(\cdot) : \mathbb{R} \to (0, 1)$ is commonly specified to be either the logit or probit link, thus obtaining $\text{pr}(y_i = 1 \mid \boldsymbol{\beta}) = [1 + \exp(-\mathbf{x}_i^\intercal \boldsymbol{\beta})]^{-1}$ in the first case, and $\text{pr}(y_i = 1 \mid \boldsymbol{\beta}) = \Phi(\mathbf{x}_i^\intercal \boldsymbol{\beta})$ in the second, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal. In performing Bayesian inference under these

models, it is common practice to specify Gaussian priors for the coefficients $\boldsymbol{\beta}$, and then update such priors with the likelihood of the observed data $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ to obtain the posterior $p(\boldsymbol{\beta} \mid \mathbf{y})$, which is used for point estimation, uncertainty quantification and prediction. However, the apparent absence of conjugacy in this Bayesian updating motivates the use of computational strategies relying either on Monte Carlo integration or on deterministic approximations (Chopin and Ridgway, 2017).

A popular class of MCMC methods that has been widely used in applications of Bayesian regression for binary data leverages augmented data representations which allow the implementation of tractable Gibbs samplers relying on conjugate full-conditional distributions. In Bayesian probit regression this strategy exploits the possibility of expressing the binary data $y_i \in \{0; 1\}$, $i = 1, \ldots, n$, as dichotomized versions of an underlying regression model for Gaussian responses $z_i \in \mathbb{R}$, $i = 1, \ldots, n$, thereby restoring conjugacy between the Gaussian prior for the coefficients $\boldsymbol{\beta}$ and the augmented data, which are in turn sampled from truncated normal full-conditionals (Albert and Chib, 1993). More recently, Polson et al. (2013) proposed a related strategy for logit regression which is based on a representation of the logistic likelihood as a scale mixture of Gaussians with respect to Pólya-gamma augmented variables $z_i \in \mathbb{R}^+$, $i = 1, \ldots, n$. Despite their simplicity, these methods face well-known computational and mixing issues in high-dimensional settings, especially with imbalanced datasets (Johndrow et al., 2019). We refer to Chopin and Ridgway (2017) for a discussion of related data-augmentation strategies (Holmes and Held, 2006; Frühwirth-Schnatter and Frühwirth, 2007) and alternative sampling methods, such as adaptive Metropolis–Hastings (Roberts and Rosenthal, 2001; Haario et al., 2001) and Hamiltonian Monte Carlo (Hoffman and Gelman, 2014), among others. While these strategies address some disadvantages of data-augmentation Gibbs samplers, they are still computationally impractical in large p applications (Chopin and Ridgway, 2017; Nishimura and Suchard, 2018; Durante, 2019).

A possible solution to scale-up computations is to consider deterministic approximations of the posterior distribution. In binary regression contexts, two strategies that have gained growing popularity are mean-field (MF) VB with global and local variables (Jaakkola and Jordan, 2000; Consonni and Marin, 2007; Durante and Rigon, 2019), and EP (Chopin and Ridgway, 2017). The first class of methods approximates the joint posterior density $p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})$ for the global parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ and the local augmented data $\mathbf{z} = (z_1, \ldots, z_n)^\intercal$ with an optimal factorized density $q^*_{\text{MF}}(\boldsymbol{\beta}) \prod_{i=1}^n q^*_{\text{MF}}(z_i)$ which is the closest in Kullback–Leibler divergence (Kullback and Leibler, 1951) to $p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})$, among all the approximating densities in the mean-field family $\mathcal{Q}_{\text{MF}} = \{q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) : q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) = q_{\text{MF}}(\boldsymbol{\beta}) q_{\text{MF}}(\mathbf{z})\}$. Optimization typically proceeds via coordinate ascent variational inference methods (CAVI) which can scale easily to large p settings. However, MF-VB is known to underestimate posterior uncertainty and often leads to Gaussian approximations which affect the quality of inference if the actual posterior is non-Gaussian (Kuss and Ras-

mussen, 2005). As we will show in Sections 2.2 and 2.3, this issue can have dramatic implications in the setting considered in this chapter. Also EP provides Gaussian approximations (Chopin and Ridgway, 2017), but typically improves the quality of VB via a moment matching of approximate marginals that have the same factorized form of the actual posterior. These gains come, however, with a computational cost which makes EP not practical for high-dimensional settings with, e.g., $p > 1000$. Indeed, recalling a concluding remark by Chopin and Ridgway (2017), the lack of scalability to large p is common to most state-of-the-art methods for Bayesian computation in binary regression models. An exception is provided by the recent contribution of Durante (2019), which proves that in Bayesian probit regression with Gaussian priors the posterior actually belongs to the class of unified skew-normal (SUN) distributions (Arellano-Valle and Azzalini, 2006). These variables have several closure properties which facilitate posterior inference in large p settings. However, the calculation of relevant functionals for inference and prediction requires the evaluation of cumulative distribution functions of n-variate Gaussians or sampling from n-variate truncated normals, thus making these results impractical in a variety of applications with sample size n greater than a few hundreds (Durante, 2019).

In this chapter we address most of the aforementioned issues by proposing a new partially factorized mean-field approximation (PFM) for Bayesian probit regression which avoids assuming independence between the global variables $\boldsymbol{\beta}$ and the augmented data $\mathbf{z}$. Unlike EP (Chopin and Ridgway, 2017), the proposed PFM-VB scales easily to $p \gg 1000$ settings, and, unlike for the computational strategies proposed in Durante (2019), it only requires evaluation of distribution functions of univariate Gaussians. Moreover, despite having a computational cost comparable to standard MF-VB for probit models (Consonni and Marin, 2007), the proposed PFM-VB leads to a substantially improved approximation of the posterior in large p settings, which reduces bias in locations and variances, and crucially incorporates skewness. Optimization proceeds via a simple CAVI algorithm and provides a tractable SUN approximating density. The methodology is discussed in Section 2.2, where we also provide theoretical results showing that the PFM-VB approximation converges to the exact posterior as $p \to \infty$, and that the number of iterations required by the CAVI to find the optimum converges to 1 as $p \to \infty$. Insightful negative results on the accuracy of standard MF-VB approximations, that suggest caution against maximum a posteriori inferences in high-dimensional contexts, are also provided. The proposed methods are evaluated on an Alzheimer's application with $p = 9036$ in Section 2.3. Concluding remarks and proofs can be found in Section 2.4 and in Appendix 2.A, respectively. Finally, Appendix 2.B discusses the computational complexity of the proposed inference and optimization strategies which can crucially be performed at an $\mathcal{O}(pn \cdot \min\{p, n\})$ cost. Codes and tutorials to implement the proposed methods and reproduce the analyses are available at `https://github.com/augustofasano/Probit-PFMVB`.

2.2 Variational Bayesian Inference for Probit Models

Recalling Section 2.1, we focus on posterior inference for the classical Bayesian probit regression model defined as

$$\begin{aligned}(y_i \mid \boldsymbol{\beta}) &\overset{ind}{\sim} \text{Bern}[\Phi(\mathbf{x}_i^\intercal \boldsymbol{\beta})], \quad i = 1, \ldots, n, \\ \boldsymbol{\beta} &\sim \text{N}_p(\mathbf{0}, \nu_p^2 \mathbf{I}_p).\end{aligned} \tag{2.1}$$

In (2.1), each y_i is a binary variable whose success probability depends on a p-dimensional vector of observed predictors $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$ under a probit mapping. The coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ regulate the effect of each predictor and are assigned independent Gaussian priors $\beta_j \sim \text{N}(0, \nu_p^2)$, for every $j = 1, \ldots, p$. Although our contribution can be naturally generalized to a generic multivariate Gaussian prior for $\boldsymbol{\beta}$, we consider here the simpler setting with $\boldsymbol{\beta} \sim \text{N}_p(\mathbf{0}, \nu_p^2 \mathbf{I}_p)$ to ease notation, and allow the prior variance ν_p^2 to possibly change with p. This choice incorporates not only routine implementations of Bayesian probit models relying on constant prior variances $\nu_p^2 = \nu^2$ for the coefficients (e.g., Chopin and Ridgway, 2017), but also more structured formulations for high-dimensional problems which define $\nu_p^2 = \nu^2/p$ to control the prior variance of the entire linear predictor and induce increasing shrinkage (e.g., Simpson et al., 2017; Fuglstad et al., 2018). The prior variance coefficient ν_p^2 is allowed to vary with p, in order to accommodate for most common routine implementations of Bayesian probit models, where either $\nu_p^2 = \nu^2$ and hence the coefficients are considered *a priori* independent with constant variance (e.g., Chopin and Ridgway, 2017) or $\nu_p^2 = \nu^2/p$ so that the variance of the linear predictor is kept constant (e.g., Fuglstad et al., 2018). Details on the technical assumptions for the asymptotic behavior of ν_p^2 are stated in Assumption A2.

Model (2.1) also has a simple constructive representation based on Gaussian augmented data, which has been broadly used in the development of MCMC (Albert and Chib, 1993) and VB (Consonni and Marin, 2007) methods. More specifically, (2.1) can be obtained by marginalizing out the augmented data $\mathbf{z} = (z_1, \ldots, z_n)^\intercal$ in the model

$$\begin{aligned}y_i &= \mathbb{1}(z_i > 0), \\ (z_i \mid \boldsymbol{\beta}) &\overset{ind}{\sim} \text{N}(\mathbf{x}_i^\intercal \boldsymbol{\beta}, 1), \quad i = 1, \ldots, n, \\ \boldsymbol{\beta} &\sim \text{N}_p(\mathbf{0}, \nu_p^2 \mathbf{I}_p).\end{aligned} \tag{2.2}$$

Recalling Albert and Chib (1993), the above construction leads to closed-form full-conditionals for $\boldsymbol{\beta}$ and $\mathbf{z}$, thus allowing the implementation of a Gibbs sampler where $p(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y}) = p(\boldsymbol{\beta} \mid \mathbf{z})$ is a Gaussian density, and each $p(z_i \mid \boldsymbol{\beta}, \mathbf{y}) = p(z_i \mid \boldsymbol{\beta}, y_i)$ is the density of a truncated normal, for $i = 1, \ldots, n$. We refer to Albert and Chib (1993) for more details regarding such a strategy. Our focus here is on large p settings where classical MCMC is often impractical, thus motivating more scalable methods relying

on approximate posteriors. In Section 2.2.1, we discuss standard MF-VB strategies for Bayesian probit models (Consonni and Marin, 2007) which rely on representation (2.2), and prove that in large p settings these approaches lead to poor approximations of the exact posterior that underestimate not only the variance but also the location, thus leading to unreliable inference and prediction. In Section 2.2.2, we address these issues via a new partially factorized variational approximation that has substantially improved practical and theoretical performance in large p settings, especially when $p \gg n$.

2.2.1 Mean-field variational Bayes

Recalling Blei et al. (2017), mean-field VB with global and local variables aims at providing a tractable approximation for the joint posterior density $p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})$ of the global parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ and the local variables $\mathbf{z} = (z_1, \ldots, z_n)^\intercal$, within the MF class of factorized densities $\mathcal{Q}_{\text{MF}} = \{q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) : q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) = q_{\text{MF}}(\boldsymbol{\beta}) q_{\text{MF}}(\mathbf{z})\}$. The optimal VB solution $q^*_{\text{MF}}(\boldsymbol{\beta}) q^*_{\text{MF}}(\mathbf{z})$ within this family is the one that minimizes the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) defined as

$$\text{KL}[q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) \,||\, p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})] = \mathbb{E}_{q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z})}[\log q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z})] - \mathbb{E}_{q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z})}[\log p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})], \tag{2.3}$$

with $q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) \in \mathcal{Q}_{\text{MF}}$. Alternatively, it is possible to obtain $q^*_{\text{MF}}(\boldsymbol{\beta}) q^*_{\text{MF}}(\mathbf{z})$ by maximizing

$$\text{ELBO}[q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z})] = \mathbb{E}_{q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z})}[\log p(\boldsymbol{\beta}, \mathbf{z}, \mathbf{y})] - \mathbb{E}_{q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z})}[\log q(\boldsymbol{\beta}, \mathbf{z})], \tag{2.4}$$

with $q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) \in \mathcal{Q}_{\text{MF}}$, since the ELBO coincides with the negative KL up to an additive constant. Recall also that the KL divergence is always non-negative and refer to Armagan and Zaretzki (2011) for the expression of $\text{ELBO}[q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z})]$ under (2.2). The maximization of (2.4) is typically easier than the minimization of (2.3), and can be performed via a simple coordinate ascent variational inference algorithm (CAVI) (e.g., Blei et al., 2017) cycling among the two steps below

$$\begin{aligned} q^{(t)}_{\text{MF}}(\boldsymbol{\beta}) &\propto \exp\{\mathbb{E}_{q^{(t-1)}_{\text{MF}}(\mathbf{z})} \log[p(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})]\}, \\ q^{(t)}_{\text{MF}}(\mathbf{z}) &\propto \exp\{\mathbb{E}_{q^{(t)}_{\text{MF}}(\boldsymbol{\beta})} \log[p(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y})]\}, \end{aligned} \tag{2.5}$$

where $q^{(t)}_{\text{MF}}(\boldsymbol{\beta})$ and $q^{(t)}_{\text{MF}}(\mathbf{z})$ are the solutions at iteration t. We refer to Blei et al. (2017) for why the updating in (2.5) iteratively optimizes the ELBO in (2.4), and highlight here how (2.5) is particularly simple to implement in Bayesian models having tractable full-conditional densities $p(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})$ and $p(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y})$. This is the case of the augmented-data representation (2.2) for the probit model in (2.1). Indeed, recalling Albert and Chib (1993) it easily follows that the full-conditionals under model (2.2) are

$$\begin{aligned} &(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y}) \sim \text{N}_p(\mathbf{V}\mathbf{X}^\intercal \mathbf{z}, \mathbf{V}), \quad \mathbf{V} = (\nu_p^{-2}\mathbf{I}_p + \mathbf{X}^\intercal\mathbf{X})^{-1}, \\ &(z_i \mid \boldsymbol{\beta}, \mathbf{y}) \overset{ind}{\sim} \begin{cases} \text{TN}[\mathbf{x}_i^\intercal \boldsymbol{\beta}, 1, (0, +\infty)], & \text{if } y_i = 1, \\ \text{TN}[\mathbf{x}_i^\intercal \boldsymbol{\beta}, 1, (-\infty, 0)], & \text{if } y_i = 0, \end{cases} \quad \text{for } i = 1, \ldots, n, \end{aligned} \tag{2.6}$$

Algorithm 1: CAVI algorithm to obtain $q^*_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) = q^*_{\text{MF}}(\boldsymbol{\beta}) \prod_{i=1}^n q^*_{\text{MF}}(z_i)$

for t *from* 1 *until convergence of* ELBO$[q^{(t)}_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z})]$ **do**

[**1**] Set

$$q^{(t)}_{\text{MF}}(\boldsymbol{\beta}) = \phi_p(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{(t)}; \mathbf{V}), \qquad \text{with } \bar{\boldsymbol{\beta}}^{(t)} = \mathbf{V}\mathbf{X}^\intercal \bar{\mathbf{z}}^{(t-1)},$$

where $\bar{\mathbf{z}}^{(t-1)}$ has elements $\bar{z}_i^{(t-1)} = \mathbf{x}_i^\intercal \bar{\boldsymbol{\beta}}^{(t-1)} + (2y_i - 1)\phi(\mathbf{x}_i^\intercal \bar{\boldsymbol{\beta}}^{(t-1)})\Phi[(2y_i - 1)\mathbf{x}_i^\intercal \bar{\boldsymbol{\beta}}^{(t-1)}]^{-1}$ for every $i = 1, \ldots, n$. In the above expression, $\phi_p(\boldsymbol{\beta} - \boldsymbol{\mu}; \boldsymbol{\Sigma})$ is the density of a generic p-variate Gaussian for $\boldsymbol{\beta}$ with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

[**2**] Set

$$q^{(t)}_{\text{MF}}(z_i) = \frac{\phi(z_i - \mathbf{x}_i^\intercal \bar{\boldsymbol{\beta}}^{(t)})}{\Phi[(2y_i - 1)\mathbf{x}_i^\intercal \bar{\boldsymbol{\beta}}^{(t)}]} \mathbb{1}[(2y_i - 1)z_i > 0]\,.$$

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

Output: $q^*_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) = q^*_{\text{MF}}(\boldsymbol{\beta}) \prod_{i=1}^n q^*_{\text{MF}}(z_i)$.

where $\mathbf{X}$ is the $n \times p$ design matrix with rows $\mathbf{x}_i^\intercal$, whereas $\text{TN}[\mu, \sigma^2, (a, b)]$ denotes a generic univariate normal distribution having mean μ, variance σ^2, and truncation to the interval (a, b). An important consequence of the conditional independence of $z_1, \ldots, z_n$ given $\boldsymbol{\beta}$ and $\mathbf{y}$, is that $q^{(t)}_{\text{MF}}(\mathbf{z}) = \prod_{i=1}^n q^{(t)}_{\text{MF}}(z_i)$ and thus the optimal MF-VB solution always factorizes as $q^*_{\text{MF}}(\boldsymbol{\beta}) q^*_{\text{MF}}(\mathbf{z}) = q^*_{\text{MF}}(\boldsymbol{\beta}) \prod_{i=1}^n q^*_{\text{MF}}(z_i)$. Replacing the densities of the above full-conditionals in the CAVI outlined in (2.5), it can be easily noted that $q^{(t)}_{\text{MF}}(\boldsymbol{\beta})$ and $q^{(t)}_{\text{MF}}(z_i)$, $i = 1, \ldots, n$, are Gaussian and truncated normal densities, respectively, with parameters as in Algorithm 1 (Consonni and Marin, 2007). Note that the actual parametric form of the optimal approximating densities follows directly from (2.5), without pre-specifying it.

Algorithm 1 relies on simple steps which basically require only updating of $\bar{\boldsymbol{\beta}}$ via matrix operations, and, unlike for EP, is computationally feasible in high-dimensional settings; see e.g., Table 2.1. Due to the Gaussian form of $q^*_{\text{MF}}(\boldsymbol{\beta})$ also the calculation of the approximate posterior moments and predictive probabilities is straightforward. The latter quantities can be easily expressed as

$$\begin{aligned} \text{pr}_{\text{MF}}(y_{\text{NEW}} = 1 \mid \mathbf{y}) &= \int \Phi(\mathbf{x}_{\text{NEW}}^\intercal \boldsymbol{\beta}) q^*_{\text{MF}}(\boldsymbol{\beta}) \text{d}\boldsymbol{\beta} \\ &= \Phi[\mathbf{x}_{\text{NEW}}^\intercal \bar{\boldsymbol{\beta}}^* (1 + \mathbf{x}_{\text{NEW}}^\intercal \mathbf{V} \mathbf{x}_{\text{NEW}})^{-1/2}], \end{aligned} \tag{2.7}$$

where $\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p$ are the covariates of the new unit, and $\bar{\boldsymbol{\beta}}^* = \mathbb{E}_{q^*_{\text{MF}}(\boldsymbol{\beta})}(\boldsymbol{\beta})$. However, as shown by the asymptotic results in Theorem 2.1, MF-VB can lead to poor approximations of the posterior in high dimensions as $p \to \infty$, causing concerns on the quality of inference and prediction. Throughout the paper, the asymptotic results are derived under the following random design assumption.

A 1. *Assume that the predictors x_{ij}, $i = 1, \ldots, n$, with $j = 1, \ldots, p$, are independent random variables with $\mathbb{E}(x_{ij}) = 0$, $\mathbb{E}(x_{ij}^2) = \sigma_x^2$ and $\sup_{ij} \mathbb{E}(x_{ij}^4) < \infty$.*

The above random design assumption is common to asymptotic studies of regression models (see e.g., Brown et al., 2002; Reiß, 2008; Qin and Hobert, 2019). Moreover, the zero mean and the constant variance assumption is a natural requirement in the context of binary regression, where the predictors are typically standardized following the recommended practice in the literature (e.g., Gelman et al., 2008; Chopin and Ridgway, 2017). In Section 2.3, we will show how empirical evidence on a real dataset, where this assumption might not hold, is still coherent with the theoretical results stated below.

To rule out pathological cases, we also require the following mild technical assumption on the behavior of ν_p^2 as $p \to \infty$.

A 2. *Assume that* $\sup_p\{\nu_p^2\} < \infty$, *and that* $\alpha = \lim_{p\to\infty} p\nu_p^2$ *exists and belongs to* $(0, \infty]$.

Observe that Assumption 2 includes the two elicitations for ν_p^2 of interest in these settings as discussed in Section 2.2 — i.e., $\nu_p^2 = \nu^2$ and $\nu_p^2 = \nu^2/p$, with $\nu^2 < \infty$. In the following, we use the convention $\alpha\sigma_x^2(1+\alpha\sigma_x^2)^{-1} = 1 = (1+\alpha\sigma_x^2)(\alpha\sigma_x^2)^{-1}$, whenever $\alpha = \infty$.

Theorem 2.1. *Under A1 and A2, we have that* $\liminf_{p\to\infty} \mathrm{KL}[q^*_{\mathrm{MF}}(\boldsymbol{\beta}) \,||\, p(\boldsymbol{\beta} \mid \mathbf{y})] > 0$ *almost surely (a.s.). Moreover,* $\nu_p^{-1}||\mathbb{E}_{q^*_{\mathrm{MF}}(\boldsymbol{\beta})}(\boldsymbol{\beta})|| \overset{a.s.}{\to} 0$ *as* $p \to \infty$, *where* $||\cdot||$ *is the usual Euclidean norm. On the contrary,* $\nu_p^{-1}||\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{y})}(\boldsymbol{\beta})|| \overset{a.s.}{\to} [\alpha\sigma_x^2(1+\alpha\sigma_x^2)^{-1}]^{1/2} c\sqrt{n} > 0$ *as* $p \to \infty$, *where* $c = 2\int_0^\infty z\phi(z)dz$ *is a strictly positive constant.*

According to Theorem 2.1, MF-VB causes over-shrinkage of the approximate posterior means, which can result in an unsatisfactory approximation of the entire posterior density $p(\boldsymbol{\beta} \mid \mathbf{y})$ in high-dimensional settings. For instance, recalling the expression of the approximate predictive probabilities in (2.7), the over-shrinkage of $\bar{\boldsymbol{\beta}}^*$ towards $\mathbf{0}$ may cause rapid concentration of $\mathrm{pr}_{\mathrm{MF}}(y_{\mathrm{NEW}} = 1 \mid \mathbf{y})$ around 0.5, thereby inducing bias. As shown in Section 2.3, the magnitude of such a bias can be dramatic, making (2.7) unreliable in high-dimensional settings. In addition, although as $p \to \infty$ the prior plays a progressively more important role in the Bayesian updating, Theorem 2.1 suggests that even few data points can induce non–negligible differences between prior and posterior moments such as, for example, the expected values.

As discussed in the proof of Theorem 2.1 and in Armagan and Zaretzki (2011), $\bar{\boldsymbol{\beta}}^*$ is also the mode of the actual posterior $p(\boldsymbol{\beta} \mid \mathbf{y})$. Hence, the above results suggest that, despite its popularity (Chopin and Ridgway, 2017; Gelman et al., 2008), the posterior mode should be avoided as a point estimate in large p settings. As a consequence, also Laplace approximation would provide unreliable inference since this approximation is centered at the posterior mode. These results are in apparent contradiction with the fact that the marginal posterior densities $p(\beta_j|\mathbf{y})$ often exhibit negligible skewness and their modes $\arg\max p(\beta_j|\mathbf{y})$ are typically close to the corresponding mean $\mathbb{E}_{p(\beta_j|\mathbf{y})}(\beta_j)$; see e.g., Figure 2.2. However, the same is not true for the joint posterior density $p(\boldsymbol{\beta}|\mathbf{y})$, where

little skewness is sufficient to induce a dramatic difference between the joint posterior mode, $\arg\max p(\boldsymbol{\beta}|\mathbf{y})$, and the posterior expectation; see e.g., Figure 2.3. In this sense, the results in Theorem 2.1 point towards caution in assessing Gaussianity of high-dimensional distributions based on the shape of their marginal distributions.

Motivated by the above considerations, in Section 2.2.2 we develop a new PFM-VB with global and local variables that solves the aforementioned issues without increasing computational costs. In fact, the cost of our procedure is the same of MF-VB but, unlike for such a strategy, we obtain a substantially improved approximation that provably converges to the exact posterior as $p \to \infty$. The magnitude of these improvements is outlined in the empirical studies in Section 2.3.

2.2.2 Partially factorized variational Bayes

A natural strategy to improve the performance of MF-VB is to relax the factorization assumptions on the approximating densities in a way that still allows simple optimization and inference. To accomplish this goal, we consider a partially factorized representation $\mathcal{Q}_{\text{PFM}} = \{q_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z}) : q_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z}) = q_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) \prod_{i=1}^{n} q_{\text{PFM}}(z_i)\}$ which does not assume independence among the parameters $\boldsymbol{\beta}$ and the local variables $\mathbf{z}$, thus providing a more flexible family of approximating densities. This new enlarged family $\mathcal{Q}_{\text{PFM}}$ allows to incorporate more structure of the actual posterior relative to $\mathcal{Q}_{\text{MF}}$, while retaining tractability. In fact, following Holmes and Held (2006) and recalling that $\mathbf{V} = (\nu_p^{-2}\mathbf{I}_p + \mathbf{X}^\intercal\mathbf{X})^{-1}$, the joint density $p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})$ under the augmented model (2.2) can be factorized as $p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y}) = p(\boldsymbol{\beta} \mid \mathbf{z})p(\mathbf{z} \mid \mathbf{y})$, where $p(\boldsymbol{\beta} \mid \mathbf{z}) = \phi_p(\boldsymbol{\beta} - \mathbf{V}\mathbf{X}^\intercal\mathbf{z}; \mathbf{V})$ and $p(\mathbf{z} \mid \mathbf{y}) \propto \phi_n(\mathbf{z}; \mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal) \prod_{i=1}^{n} \mathbb{1}[(2y_i - 1)z_i > 0]$ denote the densities of a p-variate Gaussian and an n-variate truncated normal, respectively. The main source of intractability in this factorization of the posterior is the truncated normal density, which requires the evaluation of cumulative distribution functions of n-variate Gaussians with full variance-covariance matrix for inference (Genz, 1992; Horrace, 2005; Chopin, 2011; Pakman and Paninski, 2014; Botev, 2017; Durante, 2019). The independence assumption among the augmented data in $\mathcal{Q}_{\text{PFM}}$ avoids the intractability that would arise from the multivariate truncated normal density $p(\mathbf{z} \mid \mathbf{y})$, while being fully flexible on $q_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z})$. Crucially, the optimal MF-VB approximation $q^*_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z})$ belongs to $\mathcal{Q}_{\text{PFM}}$ and thus, by minimizing $\text{KL}[q_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z}) \mid\mid p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})]$ in $\mathcal{Q}_{\text{PFM}}$, we are guaranteed to obtain an improved approximation of the joint posterior density relative to MF-VB, as stated in Proposition 2.2.

Proposition 2.2. *Let $q^*_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z})$ and $q^*_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z})$ be the optimal approximations for $p(\boldsymbol{\beta}, \mathbf{z}|\mathbf{y})$ from (2.2), under* PFM-VB *and* MF-VB*, respectively. Since $q^*_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) \in \mathcal{Q}_{\text{PFM}}$ and $q^*_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z})$ minimizes $\text{KL}[q(\boldsymbol{\beta}, \mathbf{z}) \mid\mid p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})]$ in $\mathcal{Q}_{\text{PFM}}$, then it follows $\text{KL}[q^*_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z}) \mid\mid p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})] \leq \text{KL}[q^*_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) \mid\mid p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})]$.*

This result suggests that PFM-VB may provide a promising direction to improve quality of posterior approximation. However, to be useful in practice, the solution $q^*_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z})$ should be simple to derive and the approximate posterior distribution $q^*_{\text{PFM}}(\boldsymbol{\beta}) = \int_{\mathbb{R}^n} q^*_{\text{PFM}}(\boldsymbol{\beta}|\mathbf{z}) \prod_{i=1}^n q^*_{\text{PFM}}(z_i) \mathrm{d}\mathbf{z} = \mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}[q^*_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z})]$ of direct interest should be available in tractable form. Theorem 2.3 and Corollary 2.4 show that this is possible.

Theorem 2.3. *Under the augmented model in equation* (2.2)*, the* KL *divergence between* $q_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z}) \in \mathcal{Q}_{\text{PFM}}$ *and* $p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})$ *is minimized at* $q^*_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) \prod_{i=1}^n q^*_{\text{PFM}}(z_i)$ *with*

$$\begin{aligned} q^*_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) &= p(\boldsymbol{\beta} \mid \mathbf{z}) = \phi_p(\boldsymbol{\beta} - \mathbf{V}\mathbf{X}^\intercal\mathbf{z}; \mathbf{V}), \quad \mathbf{V} = (\nu_p^{-2}\mathbf{I}_p + \mathbf{X}^\intercal\mathbf{X})^{-1}, \\ q^*_{\text{PFM}}(z_i) &= \frac{\phi(z_i - \mu_i^*; \sigma_i^{*2})}{\Phi[(2y_i - 1)\mu_i^*/\sigma_i^*]} \mathbb{1}[(2y_i - 1)z_i > 0], \quad \sigma_i^{*2} = (1 - \mathbf{x}_i^\intercal\mathbf{V}\mathbf{x}_i)^{-1}, \end{aligned} \tag{2.8}$$

for $i = 1, \ldots, n$*, where* $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_n^*)^\intercal$ *solves the system* $\mu_i^* - \sigma_i^{*2}\mathbf{x}_i^\intercal\mathbf{V}\mathbf{X}_{-i}^\intercal\bar{\mathbf{z}}_{-i}^* = 0$*,* $i = 1, \ldots, n$*, with* $\mathbf{X}_{-i}$ *denoting the design matrix without the* i*th row, while* $\bar{\mathbf{z}}_{-i}^*$ *is an* $n - 1$ *vector obtained by removing the* i*th element* $\bar{z}_i^* = \mu_i^* + (2y_i - 1)\sigma_i^*\phi(\mu_i^*/\sigma_i^*)\Phi[(2y_i - 1)\mu_i^*/\sigma_i^*]^{-1}$*,* $i = 1, \ldots, n$*, from the vector* $\bar{\mathbf{z}}^* = (\bar{z}_1^*, \ldots, \bar{z}_n^*)^\intercal$*.*

Algorithm 2: CAVI algorithm to obtain $q^*_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z}) = q^*_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) \prod_{i=1}^n q^*_{\text{MF}}(z_i)$

[**1**] Set $q^*_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) = \phi_p(\boldsymbol{\beta} - \mathbf{V}\mathbf{X}^\intercal\mathbf{z}; \mathbf{V})$ with $\mathbf{V} = (\nu_p^{-2}\mathbf{I}_p + \mathbf{X}^\intercal\mathbf{X})^{-1}$, and initialize $\mu_i^{(0)} \in \mathbb{R}$, $i = 1, \ldots, n$.

[**2**] **for** t *from* 1 *until convergence of* $\text{ELBO}[q^{(t)}_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z})]$ **do**

  for i *from* 1 *to* n **do**

    Set

$$q^{(t)}_{\text{PFM}}(z_i) = \frac{\phi(z_i - \mu_i^{(t)}; \sigma_i^{*2})}{\Phi[(2y_i - 1)\mu_i^{(t)}/\sigma_i^*]} \mathbb{1}[(2y_i - 1)z_i > 0],$$

    with $\sigma_i^{*2} = (1 - \mathbf{x}_i^\intercal\mathbf{V}\mathbf{x}_i)^{-1}$, and $\mu_i^{(t)} = \sigma_i^{*2}\mathbf{x}_i^\intercal\mathbf{V}\mathbf{X}_{-i}^\intercal(\bar{z}_1^{(t)}, \ldots, \bar{z}_{i-1}^{(t)}, \bar{z}_{i+1}^{(t-1)}, \ldots, \bar{z}_n^{(t-1)})^\intercal$ where the generic $\bar{z}_i^{(t)}$ is defined as $\bar{z}_i^*$ in Theorem 2.3 replacing μ_i^* with $\mu_i^{(t)}$.

Output: $q^*_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z}) = q^*_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) \prod_{i=1}^n q^*_{\text{PFM}}(z_i)$ and, as a consequence of Corollary 2.4, also $q^*_{\text{PFM}}(\boldsymbol{\beta})$.

In Theorem 2.3, the solution for $q^*_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z})$ follows by noting that $\text{KL}[q_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z}) \mid\mid p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})] = \text{KL}[q_{\text{PFM}}(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{y})] + \mathbb{E}_{q_{\text{PFM}}(\mathbf{z})}\{\text{KL}[q_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) \mid\mid p(\boldsymbol{\beta} \mid \mathbf{z})]\}$ due to the chain rule for the KL divergence. Thus, the second summand is 0 if and only if $q^*_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) = p(\boldsymbol{\beta} \mid \mathbf{z})$. The expressions for $q^*_{\text{PFM}}(z_i)$, $i = 1, \ldots, n$, are instead a direct consequence of the closure under conditioning property of multivariate truncated Gaussians (Horrace, 2005) which allows to recognize the kernel of a univariate truncated normal in the optimal solution $\exp[\mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z}_{-i})}(\log[p(z_i \mid \mathbf{z}_{-i}, \mathbf{y})])]$ (Blei et al., 2017) for $q^*_{\text{PFM}}(z_i)$; see Appendix 2.A for the detailed proof. Algorithm 2 outlines the steps of the CAVI to obtain $q^*_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z})$. As for classical CAVI (Blei et al., 2017), this routine optimizes the ELBO sequentially with

respect to each density $q_{\text{PFM}}(z_i)$, keeping fixed the others at their most recent update, thus producing a strategy that iteratively solves the system of equations for $\boldsymbol{\mu}^*$ in Theorem 2.3 via simple expressions. Indeed, since the form of the approximating densities is already available as in Theorem 2.3, the steps in Algorithm 2 reduce to update the vector of parameters $\boldsymbol{\mu}^*$ via simple functions and matrix operations.

As stated in Corollary 2.4, the optimal $q^*_{\text{PFM}}(\boldsymbol{\beta})$ of interest can be easily derived from $q^*_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z})$ and $\prod_{i=1}^n q^*_{\text{PFM}}(z_i)$, and coincides with the density of a tractable SUN (Arellano-Valle and Azzalini, 2006).

Corollary 2.4. *Let* $\bar{\mathbf{Y}} = \text{diag}(2y_1 - 1, \ldots, 2y_n - 1)$ *and* $\boldsymbol{\sigma}^* = \text{diag}(\sigma_1^*, \ldots, \sigma_n^*)$*, then, under* (2.8)*, the approximate density* $q^*_{\text{PFM}}(\boldsymbol{\beta})$ *for* $\boldsymbol{\beta}$ *coincides with that of the variable*

$$\mathbf{u}^{(0)} + \mathbf{V}\mathbf{X}^\intercal \bar{\mathbf{Y}} \boldsymbol{\sigma}^* \mathbf{u}^{(1)}, \tag{2.9}$$

where $\mathbf{u}^{(0)} \sim \text{N}_p(\mathbf{V}\mathbf{X}^\intercal \boldsymbol{\mu}^*, \mathbf{V})$*, and* $\mathbf{u}^{(1)} = (u_1^{(1)}, \ldots, u_n^{(1)})^\intercal$ *denotes an* n*-dimensional random vector of independent univariate truncated normals* $u_i^{(1)} \sim \text{TN}[0, 1, [-(2y_i - 1)\mu_i^*/\sigma_i^*, +\infty]]$*,* $i = 1, \ldots, n$*. Hence, recalling Arellano-Valle and Azzalini (2006) and Azzalini and Capitanio (2014),* $q^*_{\text{PFM}}(\boldsymbol{\beta})$ *is the probability density of the unified skew-normal distribution* $SUN_{p,n}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$*, with parameters*

$$\begin{aligned} \boldsymbol{\xi} &= \mathbf{V}\mathbf{X}^\intercal \boldsymbol{\mu}^*, \quad \boldsymbol{\Omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega} = \mathbf{V} + \mathbf{V}\mathbf{X}^\intercal \boldsymbol{\sigma}^{*2}\mathbf{X}\mathbf{V}, \\ \boldsymbol{\Delta} &= \boldsymbol{\omega}^{-1}\mathbf{V}\mathbf{X}^\intercal \bar{\mathbf{Y}} \boldsymbol{\sigma}^*, \quad \boldsymbol{\gamma} = \bar{\mathbf{Y}} \boldsymbol{\sigma}^{*^{-1}} \boldsymbol{\mu}^*, \quad \boldsymbol{\Gamma} = \mathbf{I}_n, \end{aligned}$$

where $\boldsymbol{\omega}$ *denotes a* $p \times p$ *diagonal matrix containing the square roots of the diagonal elements in the covariance matrix* $\boldsymbol{\Omega}$*, whereas* $\bar{\boldsymbol{\Omega}}$ *denotes the associated correlation matrix.*

The results in Corollary 2.4 follow by noticing that, under (2.8), the approximate density for $\boldsymbol{\beta}$ is the convolution of a p-variate Gaussian and an n-variate truncated normal, thereby producing the density of a SUN (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014). This class of random variables generalizes the multivariate Gaussian family via a skewness-inducing mechanism controlled by the matrix $\boldsymbol{\Delta}$ which weights the skewing effect produced by an n-variate truncated normal with covariance matrix $\boldsymbol{\Gamma}$ (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014). Besides introducing asymmetric shapes in multivariate Gaussians, the SUN has several closure properties which facilitate inference. However, the evaluation of functionals requires the calculation of cumulative distribution functions of n-variate Gaussians (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014), which is prohibitive when n is large, unless $\boldsymbol{\Gamma}$ is diagonal. Recalling Durante (2019), this issue makes Bayesian inference rapidly impractical under the exact posterior $p(\boldsymbol{\beta} \mid \mathbf{y})$ when n is more than a few hundreds, since $p(\boldsymbol{\beta} \mid \mathbf{y})$ is a SUN density with non-diagonal $\boldsymbol{\Gamma}_{\text{post}}$. Instead, the factorized form $\prod_{i=1}^n q_{\text{PFM}}(z_i)$ for

$q_{\text{PFM}}(\mathbf{z})$ leads to a SUN approximate density for $\boldsymbol{\beta}$ in Corollary 2.4, which crucially relies on a diagonal $\boldsymbol{\Gamma} = \mathbf{I}_n$. Such a result allows approximate posterior inference for every n and p via tractable expressions. In particular, recalling the stochastic representation in (2.9), the first two central moments of $\boldsymbol{\beta}$ and the predictive distribution are derived in Proposition 2.5.

Proposition 2.5. *If $q^*_{\text{PFM}}(\boldsymbol{\beta})$ is the SUN density in Corollary 2.4, then*

$$\begin{aligned} \mathbb{E}_{q^*_{\text{PFM}}(\boldsymbol{\beta})}(\boldsymbol{\beta}) &= \mathbf{V}\mathbf{X}^\intercal \bar{\mathbf{z}}^*, \\ \text{var}_{q^*_{\text{PFM}}(\boldsymbol{\beta})}(\boldsymbol{\beta}) &= \mathbf{V} + \mathbf{V}\mathbf{X}^\intercal \mathbf{C}^* \mathbf{X}\mathbf{V}, \end{aligned} \tag{2.10}$$

where $\mathbf{C}^* = \text{diag}[\sigma_1^{*2} - (\bar{z}_1^* - \mu_1^*)\bar{z}_1^*, \ldots, \sigma_n^{*2} - (\bar{z}_n^* - \mu_n^*)\bar{z}_n^*]$, and $\bar{z}_i^*$, μ_i^* and σ_i^*, $i = 1, \ldots, n$ are defined as in Theorem 2.3. Moreover, the posterior predictive probability $\text{pr}_{\text{PFM}}(y_{\text{NEW}} = 1 \mid \mathbf{y}) = \int \Phi(\mathbf{x}_{\text{NEW}}^\intercal \boldsymbol{\beta}) q^*_{\text{PFM}}(\boldsymbol{\beta}) \text{d}\boldsymbol{\beta}$ for a new unit with covariates $\mathbf{x}_{\text{NEW}}$ is

$$\text{pr}_{\text{PFM}}(y_{\text{NEW}} = 1 \mid \mathbf{y}) = \mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}\{\Phi[\mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{X}^\intercal \mathbf{z}(1 + \mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{x}_{\text{NEW}})^{-1/2}]\}, \tag{2.11}$$

where, according to Theorem 2.3, $q^*_{\text{PFM}}(\mathbf{z})$ can be expressed as the product $\prod_{i=1}^n q^*_{\text{PFM}}(z_i)$ of univariate truncated normal densities $q^*_{\text{PFM}}(z_i) = \phi(z_i - \mu_i^*; \sigma_i^{*2})\Phi[(2y_i - 1)\mu_i^*/\sigma_i^*]^{-1}\mathbb{1}[(2y_i - 1)z_i > 0]$, $i = 1, \ldots, n$.

Hence, unlike for inference under the exact posterior (Durante, 2019), calculation of relevant approximate moments such as those in equation (2.10), only requires the evaluation of cumulative distribution functions of univariate Gaussians. Similarly, the predictive probabilities in equation (2.11) can be easily evaluated via efficient Monte Carlo methods based on samples from n independent univariate truncated normals with density $q^*_{\text{PFM}}(z_i)$, $i = 1, \ldots, n$. Moreover, leveraging (2.9), samples from the approximate posterior $q^*_{\text{PFM}}(\boldsymbol{\beta})$ can directly be obtained via a linear combination between realizations from a p-variate Gaussian and from n univariate truncated normals, as shown in Algorithm 3. This strategy allows to study complex approximate functionals of $\boldsymbol{\beta}$ through simple Monte Carlo methods. If instead the focus is only on $q^*_{\text{PFM}}(\beta_j)$, $j = 1, \ldots, p$, one can avoid the cost of simulating from the p-variate Gaussian in Algorithm 3 and just sample from the marginals of $\mathbf{u}^{(0)}$ in the additive representation of the SUN to get samples from $q^*_{\text{PFM}}(\beta_j)$ for $j = 1, \ldots, p$ at an $\mathcal{O}(pn \cdot \min\{p, n\})$ cost.

Algorithm 3: Strategy to sample from the approximate SUN posterior in Corollary 2.4

[**1**] Draw $\mathbf{u}^{(0)} \sim \text{N}_p(\mathbf{V}\mathbf{X}^\intercal \boldsymbol{\mu}^*, \mathbf{V})$.
[**2**] Draw $u_i^{(1)} \sim \text{TN}[0, 1, [-(2y_i - 1)\mu_i^*/\sigma_i^*, +\infty]]$, $i = 1, \ldots, n$. Set $\mathbf{u}^{(1)} = (u_1^{(1)}, \ldots, u_n^{(1)})^\intercal$.
[**3**] Compute $\boldsymbol{\beta} = \mathbf{u}^{(0)} + \mathbf{V}\mathbf{X}^\intercal \bar{\mathbf{Y}} \boldsymbol{\sigma}^* \mathbf{u}^{(1)}$.

Output: a draw $\boldsymbol{\beta}$ from the approximate posterior with density as in (2.9).

We conclude the presentation of PFM-VB by studying its properties in high-dimensional settings as $p \to \infty$. As discussed in Section 2.2.1, MF-VB (Consonni and Marin, 2007) provides poor Gaussian approximations of the posterior density in high dimensions, which do not include asymmetric shapes usually found in Bayesian binary regression (Kuss and Rasmussen, 2005), and affect quality of inference and prediction. By relaxing the MF assumption we obtain, instead, an approximate density which includes skewness and matches the exact posterior for $\boldsymbol{\beta}$ when $p \to \infty$, as stated in Theorem 2.6.

Theorem 2.6. *Under A1 and A2, we have that* $\text{KL}[q^*_{\text{PFM}}(\boldsymbol{\beta}) \,||\, p(\boldsymbol{\beta} \mid \mathbf{y})] \xrightarrow{a.s.} 0$ *as* $p \to \infty$.

Hence, in the high dimensional settings where current computational strategies are impractical (Chopin and Ridgway, 2017), inference and prediction under the approximation provided by PFM-VB is practically feasible, and provides essentially the same results as those obtained under the exact posterior. For instance, Corollary 2.7 states that, unlike for MF-VB, PFM-VB is guaranteed to provide increasingly accurate approximations of posterior predictive probabilities as $p \to \infty$.

Corollary 2.7. *Let* $\text{pr}(y_{\text{NEW}} = 1 \mid \mathbf{y}) = \int \Phi(\mathbf{x}^\intercal_{\text{NEW}}\boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{y}) \text{d}\boldsymbol{\beta}$ *denote the exact posterior predictive probability for a new observation with predictors* $\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p$*, then, under A1 and A2, we have that* $\sup_{\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p} |\text{pr}_{\text{PFM}}(y_{\text{NEW}} = 1 \mid \mathbf{y}) - \text{pr}(y_{\text{NEW}} = 1 \mid \mathbf{y})| \xrightarrow{a.s.} 0$ *as* $p \to \infty$*. On the contrary,* $\liminf_{p\to\infty} \sup_{\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p} |\text{pr}_{\text{MF}}(y_{\text{NEW}} = 1 \mid \mathbf{y}) - \text{pr}(y_{\text{NEW}} = 1 \mid \mathbf{y})| > 0$ *almost surely as* $p \to \infty$.

Corollary 2.7 implies that, under A1 and A2, the error made by PFM-VB in terms of approximation of posterior predictive probabilities goes to 0 as $p \to \infty$, regardless of the choice of $\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p$. On the contrary, under MF-VB there always exists, for every p, some $\mathbf{x}_{\text{NEW}}$ such that the corresponding posterior predictive probability is not accurately approximated.

Finally, as stated in Theorem 2.8, the number of iterations required by the CAVI in Algorithm 2 to produce the optimal solution $q^*_{\text{PFM}}(\boldsymbol{\beta})$ converges to 1 as $p \to \infty$.

Theorem 2.8. *Let* $q^{(t)}_{\text{PFM}}(\boldsymbol{\beta}) = \int_{\mathbb{R}^n} q^{(t)}_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) \prod_{i=1}^n q^{(t)}_{\text{PFM}}(z_i) \text{d}\mathbf{z}$ *denote the approximate density for* $\boldsymbol{\beta}$ *produced at iteration* t *by Algorithm 2. Then, under A1 and A2,* $\text{KL}[q^{(1)}_{\text{PFM}}(\boldsymbol{\beta}) \,||\, p(\boldsymbol{\beta} \mid \mathbf{y})] \xrightarrow{a.s.} 0$ *as* $p \to \infty$.

According to Theorem 2.8, the CAVI in Algorithm 2 converges essentially in one iteration as $p \to \infty$. Thus the computational complexity of the entire PFM-VB routine is provably equal to that of a single CAVI iteration, which is dominated by an $\mathcal{O}(pn \cdot \min\{p, n\})$ pre-computation cost discussed in detail in Appendix 2.B, where we also highlight how the calculation of the functionals in Proposition 2.5 can be achieved at the same cost. More complex functionals of the joint approximate posterior can be instead obtained at higher costs via Monte Carlo methods based on Algorithm 3.

Finally, we shall emphasize that also the computational complexity of approximate inference under MF-VB is dominated by the same $\mathcal{O}(pn \cdot \min\{p, n\})$ pre-computation cost. However, as discussed in Sections 2.2.1 and 2.2.2, PFM-VB produces substantially more accurate approximations relative to MF-VB. In Section 2.3, we provide further evidence for these arguments and discuss how the theoretical results presented in Sections 2.2.1 and 2.2.2 match closely the empirical behavior observed in a real-world application to Alzheimer's data.

2.3 High-Dimensional Probit Regression Application to Alzheimer's Data

As shown in Chopin and Ridgway (2017), state-of-the-art computational methods for Bayesian binary regression, such as Hamiltonian Monte Carlo (Hoffman and Gelman, 2014), VB (Consonni and Marin, 2007) and EP (Chopin and Ridgway, 2017) are feasible and powerful procedures in small-to-moderate p settings, but become rapidly impractical or inaccurate in large p contexts, such as $p > 1000$. The overarching focus of the present chapter is to close this gap and, consistent with this aim, we consider a large p study to quantify the drawbacks encountered by the aforementioned strategies along with the improvements provided by the proposed PFM-VB method.

Following the above remarks, we focus on an application to model presence or absence of Alzheimer's disease in its early stages as a function of demographic data, genotype and assay results. The original dataset is available in the `R` library `AppliedPredictiveModeling` and arises from a study of the Washington University to determine if biological measurements from cerebrospinal fluid are useful in modeling and predicting early stages of Alzheimer's disease (Craig-Schapiro et al., 2011). In the original chapter, the authors consider a variety of machine learning procedures to improve the flexibility relative to a basic binary regression model. Here, we avoid excessively complex black-box algorithms and rely on an interpretable probit regression (2.1), which improves flexibility by simply adding pairwise interactions, thus obtaining $p = 9036$ predictors collected for 333 individuals. Following Gelman et al. (2008) and Chopin and Ridgway (2017) the original measurements have been standardized to have mean 0 and standard deviation 0.5, before entering such variables and their interactions in the probit regression. In general, we recommend to always standardize the predictors when implementing PFM-VB since this choice typically reduces the correlation between units and thus also between the associated latent variables z_i, making the resulting variational approximation more accurate. We shall also emphasize that the sample size of this study is low relative to those that can be easily handled under PFM-VB. In fact, this moderate n is required to make inference under the exact posterior, which serves here as a benchmark, still feasible (Durante,

2019).

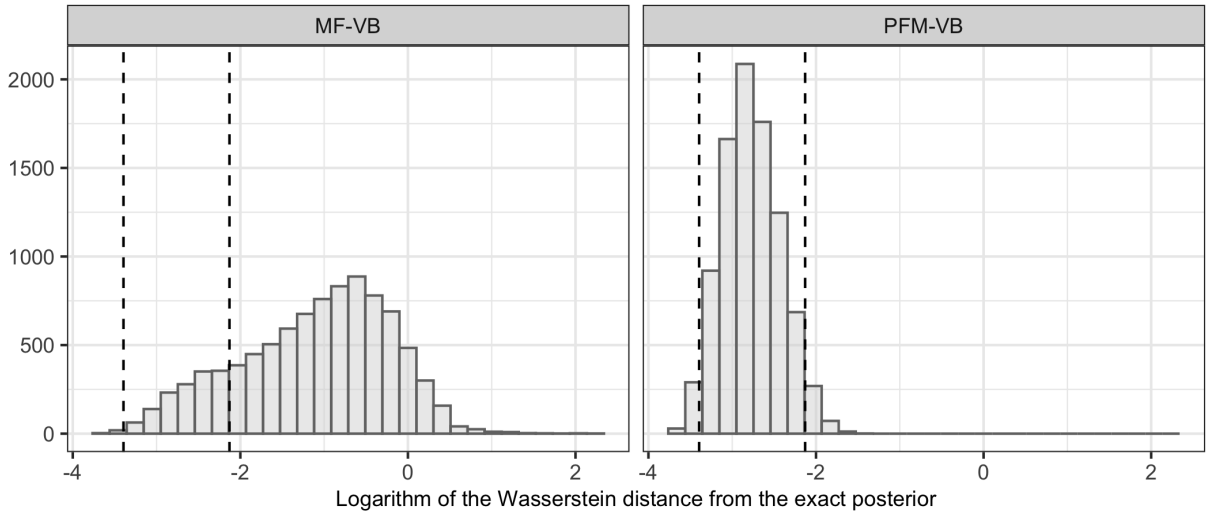

Figure 2.1: For MF-VB and PFM-VB, histograms of the log-Wasserstein distances between the $p = 9036$ approximate marginal densities provided by the two VB methods and the exact posterior marginals. These distances are computed via Monte Carlo based on 20000 samples from the approximate and exact marginals. To provide insights on Monte Carlo error, the dashed vertical lines represent the quantiles 2.5% and 97.5% of the log-Wasserstein distances between two different samples of 20000 draws from the exact posterior marginals.

In performing Bayesian inference under the above probit model, we follow the guidelines in Gelman et al. (2008) and rely on independent weakly informative Gaussian priors with mean 0 and standard deviation 5 for each coefficient β_j, $j = 1, \dots, 9036$. These priors are then updated with the likelihood of $n = 300$ units, after holding out 33 individuals to study the behavior of the posterior predictive probabilities in such large p settings, along with the performance of the overall approximation of the posterior. Table 2.1 provides insights on the computational time of MF-VB and PFM-VB, and highlights the bottlenecks encountered by relevant routine-use competitors. These include the `rstan` implementation of Hamiltonian Monte Carlo, the EP algorithm in the `R` library `EPGLM`, and the Monte Carlo strategy based on 20000 independent draws from the exact SUN posterior using the algorithm in Durante (2019). As expected, these strategies are clearly impractical in such settings. In particular, STAN and EP suffer from the large p, whereas sampling from the exact posterior is still feasible, but requires a non-negligible computational effort due to the moderately large n. Variational inference under MF-VB and PFM-VB is orders of magnitude faster and, hence, provides the only viable approach in such settings. These results motivate our main focus on the quality of MF-VB and PFM-VB approximations in Figures 2.1–2.3, taking as benchmark Monte Carlo inference based on 20000 independent samples from the exact SUN posterior. In this example PFM-VB requires only 7 CAVI iterations to converge, instead of 212 as for MF-VB. This result is in line with Theorem 2.8, and with the subsequent considerations.

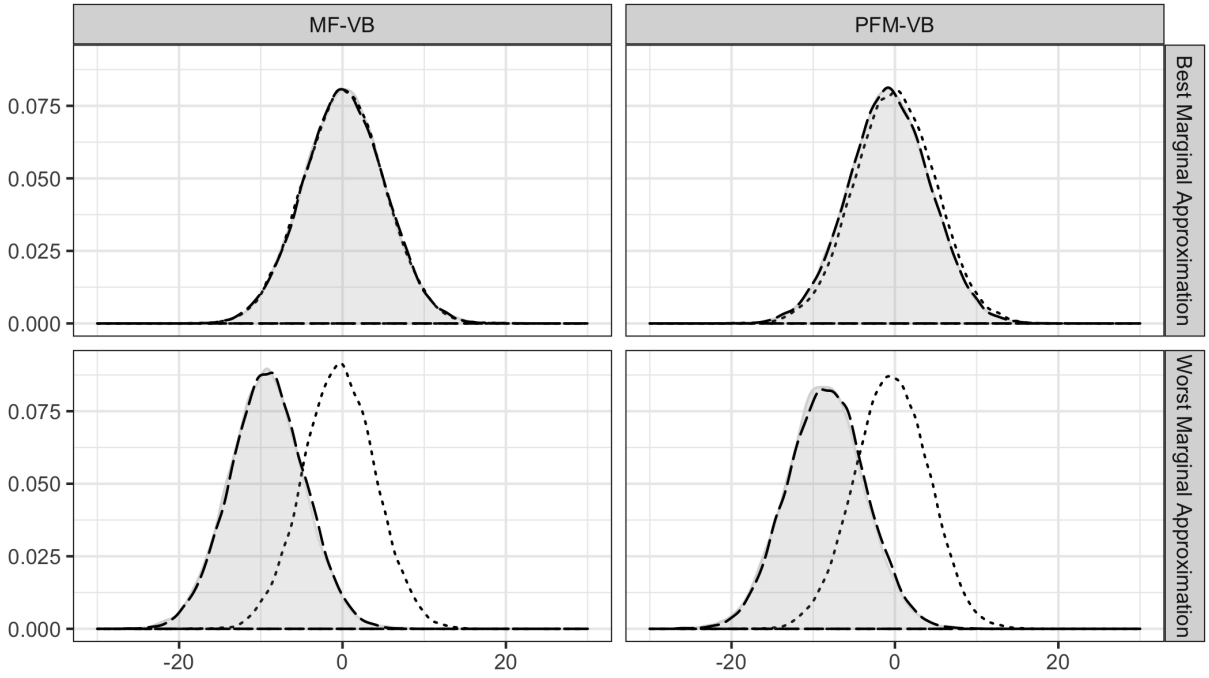

Figure 2.2: Quality of marginal approximation for the coefficients associated with the highest and lowest Wasserstein distance from the exact posterior under MF-VB and PFM-VB, respectively. The shaded grey area denotes the density of the exact posterior marginal, whereas the dotted and dashed lines represent the approximate densities provided by MF-VB and PFM-VB, respectively.

Table 2.1: Computational time of state-of-the-art routines in the Alzheimer's application. This includes the running time of the sampling or optimization procedure and the time to compute means, standard deviations and predictive probabilities, for those routines that were feasible.

	STAN	EP	SUN	MF-VB	PFM-VB
Running time in minutes	> 360.00	> 360.00	92.71	0.05	0.05

Figure 2.1 shows the histograms of the log-Wasserstein distances among the $p = 9036$ exact posterior marginals and the associated approximations under MF-VB and PFM-VB. Such quantities are computed with the `R` function `wasserstein1d`, which uses 20000 values sampled from the approximate and exact marginals. According to these histograms, PFM-VB improves the quality of MF-VB and, in practice, it matches almost perfectly the exact posterior since it provides distances within the range of values obtained by comparing two different samples of 20000 draws from the same exact posterior marginals. Hence, most of the variability in the PFM-VB histogram is arguably due to Monte Carlo error.

These results are in line with Theorems 2.1 and 2.6, and are also confirmed by Figure 2.2 which compares graphically the quality of the marginal approximation for the coefficients associated with the highest and lowest Wasserstein distance from the exact posterior

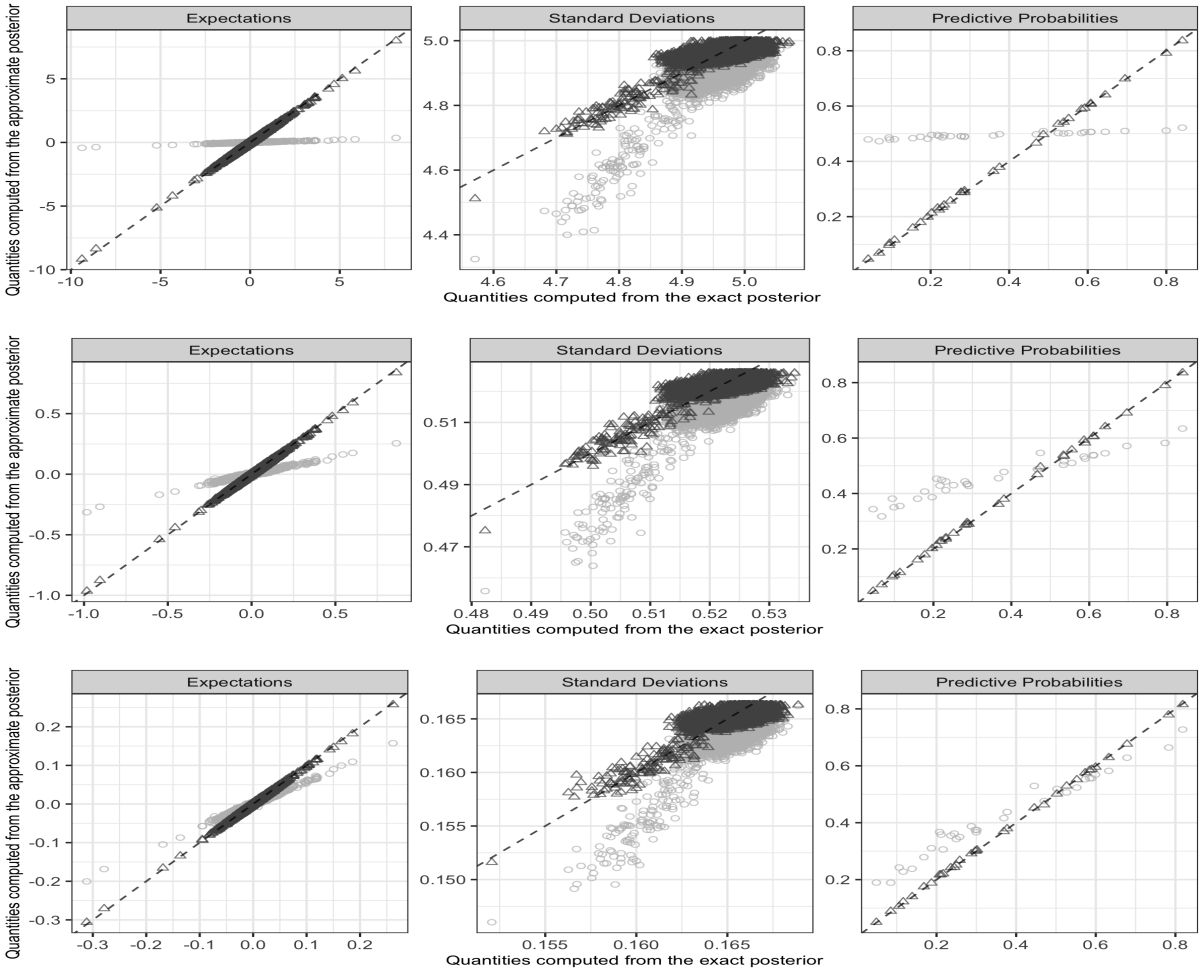

Figure 2.3: Scatterplots comparing the posterior expectations, standard deviations and predictive probabilities computed from 20000 values sampled from the exact SUN posterior, with those provided by the MF-VB (light grey circles) and PFM-VB (dark grey triangles). Each row represents a different scenario, respectively $\nu_p^2 = 25, 25 \cdot 100/p, 25 \cdot 10/p$.

under MF-VB and PFM-VB. As is clear from Figure 2.2, PFM-VB produces approximations which perfectly overlaps with the exact posterior in all cases, including also the worst-case scenario with the highest Wasserstein distance. Consistent with Theorem 2.1, MF-VB has instead a reduced quality mostly due to a tendency to shrink, sometimes dramatically, towards zero the locations of the actual posterior. This behavior is studied more in detail in Figure 2.3, where posterior expectations and standard deviations are shown, together with the predictive probabilities for the held-out observations, for the scenarios $\nu_p^2 = 25$ considered so far and also for $\nu_p^2 = 25 \cdot 100/p$ and $\nu_p^2 = 25 \cdot 10/p$. These two last values for ν_p^2 correspond to fixing the total variance of the linear predictor as if there were respectively 100 and 10 coefficients with prior standard deviation 5, in line with Gelman et al. (2008), while the others were fixed to zero. The over-shrinkage of the posterior means can be seen in the first column of Figure 2.3, which compares the posterior expectations computed from 20000 values sampled from the exact SUN posterior with those provided

by the closed-form expressions under MF-VB and PFM-VB reported in Section 2.2. We can notice that such a behavior is dramatic for the case of constant prior variance, and still remains significant when ν_p^2 is allowed to decrease with p. Also the standard deviations are slightly under-estimated relative to PFM-VB that notably removes bias also in the second order moments. Consistent with the results in Figures 2.1–2.2, the slight variability of the PFM-VB estimates in the second column of Figure 2.3 is arguably due to Monte Carlo error. We conclude by assessing quality in the approximation of the exact posterior predictive probabilities for the 33 held-out individuals. These measures are fundamental for prediction and, unlike for the first two marginal moments, their evaluation depends on the behavior of the entire posterior since it relies on a non-linear mapping of a linear combination of the parameters $\boldsymbol{\beta}$. In the third column of Figure 2.3, the proposed PFM-VB essentially matches the exact posterior predictive probabilities, thus providing reliable classification and uncertainty quantification. Instead, as expected from the theoretical results in Corollary 2.7, MF-VB over-shrinks these quantities towards 0.5.

2.4 Discussion and Future Research Directions

This chapter highlights notable issues in state-of-the-art methods for approximate Bayesian inference in high-dimensional binary regression, and proposes a partially factorized mean-field variational Bayes strategy which provably covers these open gaps. Our basic idea is to relax the mean-field assumption in a way which approximates more closely the factorization of the actual posterior, but still allows simple optimization and inference. The theoretical results confirm that the proposed strategy is an optimal solution in large p settings, especially when $p \gg n$, and the empirical studies suggest that the theory provides useful insights also in applications not necessarily meeting the assumptions.

While our contribution provides an important advancement in a non-Gaussian regression context where previously available Bayesian computational strategies are unsatisfactory (Chopin and Ridgway, 2017), the results in this chapter open new avenues for future research. For instance, the theoretical issues of MF-VB and MAP estimators presented in Section 2.2.1 for large p settings point to the need of further theoretical studies on the use of MF-VB and MAP estimators in high-dimensional regression with non-Gaussian responses. In these contexts, our general idea of relying on a partially factorized approximating family could provide a viable strategy to solve potential issues of current approximations, as long as simple optimization is possible and the approximate posterior density for the global parameters can be derived in closed-form via marginalization of the local variables. This strategy could be also useful in Bayesian models relying on hierarchical priors for $\boldsymbol{\beta}$ that facilitate variable selection and improved shrinkage. Albeit interesting, this setting goes beyond the scope of the contribution.

Finally, it would be certainly relevant to extend the asymptotic results in Theorems 2.1, 2.6 and 2.8 to settings in which n grows with p at some rate. In particular, we conjecture that n growing sublinearly with p is a sufficient condition to obtain asymptotic-exactness results analogous to Theorem 2.6.

2.A Appendix: Proofs

We start by proving some general lemmas that will be useful for the proofs of Theorems 2.1, 2.6 and 2.8. A key one is a variant of the strong law of large numbers, which is a classical result that follows from Khintchine–Kolmogorov convergence theorem and Kronecker's lemma. In the following, when we use the notation $o(p^d)$ in a matrix context, we indicate a matrix whose entries are all $o(p^d)$.

Lemma 2.9. *Let $(w_j)_{j\geq 1}$ be a sequence of independent random variables with mean 0 and variance bounded over j. Then $p^{-1/2-\delta}\sum_{j=1}^{p} w_j \overset{a.s.}{\to} 0$ as $p \to \infty$ for every $\delta > 0$.*

Lemma 2.10. *Under A1 and A2, for any $\delta > 0$ we have $(\sigma_x^2 p)^{-1}\mathbf{X}\mathbf{X}^\intercal \overset{a.s.}{=} \mathbf{I}_n + o(p^{-1/2+\delta})$ and $(1+\sigma_x^2 p\nu_p^2)^{-1}(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal) \overset{a.s.}{\to} \mathbf{I}_n$ as $p \to \infty$.*

Proof. By A1, $(x_{ij}^2)_{j\geq 1}$ are independent random variables with mean σ_x^2 and variance bounded over j. Thus, $p^{1/2-\delta}[(\sigma_x^2 p)^{-1}\mathbf{X}\mathbf{X}^\intercal - \mathbf{I}_n]_{ii} = p^{-1/2-\delta}\sum_{j=1}^{p}(\sigma_x^{-2}x_{ij}^2 - 1) \overset{a.s.}{\to} 0$ by Lemma 2.9. Similarly, when $i \neq i'$, $(x_{ij}x_{i'j})_{j\geq 1}$ are independent random variables with mean 0 and variance $\sigma_x^4 < \infty$. Thus

$$p^{1/2-\delta}[(\sigma_x^2 p)^{-1}\mathbf{X}\mathbf{X}^\intercal - \mathbf{I}_n]_{ii'} = \sigma_x^{-2}p^{-1/2-\delta}\sum_{j=1}^{p} x_{ij}x_{i'j} \overset{a.s.}{\to} 0$$

as $p \to \infty$ by Lemma 2.9. It follows that $(\sigma_x^2 p)^{-1}\mathbf{X}\mathbf{X}^\intercal \overset{a.s.}{=} \mathbf{I}_n + o(p^{-1/2+\delta})$ as $p \to \infty$. Finally, since by A2 $p\nu_p^2$ converges to a positive constant or goes to infinity, in both cases we have that, as $p \to \infty$,

$$(1+\sigma_x^2 p\nu_p^2)^{-1}(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal) = (1+\sigma_x^2 p\nu_p^2)^{-1}\mathbf{I}_n + (1+\sigma_x^2 p\nu_p^2)^{-1}(\sigma_x^2 p\nu_p^2)(\sigma_x^2 p)^{-1}\mathbf{X}\mathbf{X}^\intercal \overset{a.s.}{\to} \mathbf{I}_n.$$

□

Lemma 2.11. *Let $\mathbf{H} = \mathbf{X}\mathbf{V}\mathbf{X}^\intercal$, then, under A1 and A2, it holds $(1+\sigma_x^2 p\nu_p^2)(\mathbf{I}_n - \mathbf{H}) \overset{a.s.}{\to} \mathbf{I}_n$ as $p \to \infty$. In particular, for $p \to \infty$, we have $\mathbf{H} \overset{a.s.}{\to} \frac{\alpha\sigma_x^2}{1+\alpha\sigma_x^2}\mathbf{I}_n$ when $\alpha \in (0,\infty)$, while $\mathbf{H} \overset{a.s.}{=} [1-(\sigma_x^2 p\nu_p^2)^{-1}]\mathbf{I}_n + o(p^{-1})$ when $\alpha = \infty$.*

Proof. Since $\mathbf{V} = (\nu_p^{-2}\mathbf{I}_p + \mathbf{X}^\intercal\mathbf{X})^{-1}$, by applying the Woodbury's identity to $(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal)^{-1}$, we obtain $(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal)^{-1} = \mathbf{I}_n - \nu_p^2\mathbf{X}(\mathbf{I}_p + \nu_p^2\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal = \mathbf{I}_n - \mathbf{X}(\nu_p^{-2}\mathbf{I}_p + \mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal = \mathbf{I}_n - \mathbf{H}$. Thus $[(1+\sigma_x^2 p\nu_p^2)(\mathbf{I}_n - \mathbf{H})]^{-1} = (1+\sigma_x^2 p\nu_p^2)^{-1}(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal) \overset{a.s.}{\to} \mathbf{I}_n$ as $p \to \infty$ by Lemma 2.10 and the thesis follows by the continuity of the inverse operator over the set of non-singular $n \times n$ matrices. □

Lemma 2.12. *Let* $\boldsymbol{\mu}_l^{(p)} \to \mathbf{0}$ *and* $\boldsymbol{\Sigma}_l^{(p)} \to \mathbf{I}_n$ *as* $p \to \infty$ *for* $l = 1, 2$*, where* $\boldsymbol{\mu}_l^{(p)} \in \mathbb{R}^n$ *and* $\boldsymbol{\Sigma}_l^{(p)} \in \mathbb{R}^{n^2}$*,* $l = 1, 2$*. Then* $\text{KL}[\text{TN}(\boldsymbol{\mu}_1^{(p)}, \boldsymbol{\Sigma}_1^{(p)}, \mathbb{A}) \,||\, \text{TN}(\boldsymbol{\mu}_2^{(p)}, \boldsymbol{\Sigma}_2^{(p)}, \mathbb{A})] \to 0$ *as* $p \to \infty$*, where* $\mathbb{A}$ *is an orthant of* $\mathbb{R}^n$*.*

Proof. By definition, $\text{KL}[\text{TN}(\boldsymbol{\mu}_1^{(p)}, \boldsymbol{\Sigma}_1^{(p)}, \mathbb{A}) \,||\, \text{TN}(\boldsymbol{\mu}_2^{(p)}, \boldsymbol{\Sigma}_2^{(p)}, \mathbb{A})]$ is equal to

$$\log[(\psi_1^{(p)})^{-1}\psi_2^{(p)}] + \frac{1}{2}\log[\det(\boldsymbol{\Sigma}_1^{(p)})^{-1}\det(\boldsymbol{\Sigma}_2^{(p)})] \\ +(\psi_1^{(p)}2\pi)^{-n/2}\det(\boldsymbol{\Sigma}_1^{(p)})^{-1/2}\int_{\mathbb{A}} f_p(\mathbf{u})d\mathbf{u}\,,$$

where $\psi_l^{(p)} = \text{pr}(\mathbf{u}_l^{(p)} \in \mathbb{A})$ with $\mathbf{u}_l^{(p)} \sim \text{N}_n(\boldsymbol{\mu}_l^{(p)}, \boldsymbol{\Sigma}_l^{(p)})$, for $l = 1, 2$, and

$$f_p(\mathbf{u}) = g_p(\mathbf{u})\exp[-0.5(\mathbf{u} - \boldsymbol{\mu}_1^{(p)})^\intercal(\boldsymbol{\Sigma}_1^{(p)})^{-1}(\mathbf{u} - \boldsymbol{\mu}_1^{(p)})], \quad \text{with}$$
$$g_p(\mathbf{u}) = -0.5[(\mathbf{u} - \boldsymbol{\mu}_1^{(p)})^\intercal(\boldsymbol{\Sigma}_1^{(p)})^{-1}(\mathbf{u} - \boldsymbol{\mu}_1^{(p)}) - (\mathbf{u} - \boldsymbol{\mu}_2^{(p)})^\intercal(\boldsymbol{\Sigma}_2^{(p)})^{-1}(\mathbf{u} - \boldsymbol{\mu}_2^{(p)})].$$

Since $\boldsymbol{\mu}_l^{(p)} \to \mathbf{0}$ and $\boldsymbol{\Sigma}_l^{(p)} \to \mathbf{I}_n$ as $p \to \infty$, we have that $\text{N}_n(\boldsymbol{\mu}_l^{(p)}, \boldsymbol{\Sigma}_l^{(p)}) \to \text{N}_n(\mathbf{0}, \mathbf{I}_n)$ in distribution and $\psi_l^{(p)} \to 2^{-n}$ by Portmanteau theorem, which implies $\log[(\psi_1^{(p)})^{-1}\psi_2^{(p)}] \to 0$. In addition, by the continuity of $\det(\cdot)$, we have $\det(\boldsymbol{\Sigma}_l^{(p)}) \to \det(\mathbf{I}_n) = 1$ as $p \to \infty$, and thus

$$\log[\det(\boldsymbol{\Sigma}_1^{(p)})^{-1}\det(\boldsymbol{\Sigma}_2^{(p)})] \to 0$$

as $p \to \infty$. Moreover, $\boldsymbol{\Sigma}_l^{(p)} \to \mathbf{I}_n$ implies that all the eigenvalues of $\boldsymbol{\Sigma}_l^{(p)}$ converge to 1 as $p \to \infty$ for $l = 1, 2$, and thus are eventually bounded away from 0 and ∞. Therefore, there exist positive, finite constants m, M and k such that $m\|\mathbf{u} - \boldsymbol{\mu}_l^{(p)}\|^2 \leq (\mathbf{u} - \boldsymbol{\mu}_l^{(p)})^\intercal(\boldsymbol{\Sigma}_l^{(p)})^{-1}(\mathbf{u} - \boldsymbol{\mu}_l^{(p)}) \leq M\|\mathbf{u} - \boldsymbol{\mu}_l^{(p)}\|^2$ for $l = 1, 2$ and $p \geq k$. Calling $b = \sup_{p\geq 1,\ l\in\{1,2\}} \|\boldsymbol{\mu}_l^{(p)}\| < \infty$, and using standard properties of norms, we obtain, for $l = 1, 2$ and $p \geq k$,

$$m(\|\mathbf{u}\|^2 - 2b\|\mathbf{u}\|) \leq (\mathbf{u} - \boldsymbol{\mu}_l^{(p)})^\intercal(\boldsymbol{\Sigma}_l^{(p)})^{-1}(\mathbf{u} - \boldsymbol{\mu}_l^{(p)}) \leq (M\|\mathbf{u}\|^2 + Mb),$$

from which we immediately obtain that, for $p \geq k$, $|f_p(\mathbf{u})| \leq (M\|\mathbf{u}\|^2 + Mb)\exp(-m\|\mathbf{u}\|^2/2 + b\|\mathbf{u}\|)$, where the latter is an integrable function on $\mathbb{R}^n$. Therefore we can apply the dominated convergence theorem and obtain $\lim_{p\to\infty}\int_{\mathbb{A}} f_p(\mathbf{u})d\mathbf{u} = \int_{\mathbb{A}}\lim_{p\to\infty} f_p(\mathbf{u})d\mathbf{u} = 0$ as desired. □

2.A.1 Proof of Theorem 2.1

Let $\bar{\boldsymbol{\beta}}^* = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \ell(\boldsymbol{\beta})$, where $\ell(\boldsymbol{\beta}) = -(2\nu_p^2)^{-1}\|\boldsymbol{\beta}\|^2 + \sum_{i=1}^n \log\Phi[(2y_i - 1)\mathbf{x}_i^\intercal\boldsymbol{\beta}]$ denotes the log-posterior up to an additive constant under (2.1). Note that $\bar{\boldsymbol{\beta}}^*$ is unique because $\ell(\boldsymbol{\beta})$ is strictly concave (Haberman, 1974).

Lemma 2.13. *Under A1, we have* $\nu_p^{-1}\|\bar{\boldsymbol{\beta}}^*\| \xrightarrow{a.s.} 0$ *as* $p \to \infty$*.*

Proof. Since $\log \Phi[(2y_i-1)\mathbf{x}_i^\intercal \boldsymbol{\beta}] < 0$ for every $i = 1, \dots, n$, we obtain $\ell(\boldsymbol{\beta}) < -(2\nu_p^2)^{-1}\|\boldsymbol{\beta}\|^2$ and thus $\nu_p^{-2}\|\boldsymbol{\beta}\|^2 < -0.5\ell(\boldsymbol{\beta})$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$. It follows that $\nu_p^{-2}\|\bar{\boldsymbol{\beta}}^*\|^2 < -0.5\ell(\bar{\boldsymbol{\beta}}^*) = -0.5 \sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \ell(\boldsymbol{\beta})$. We now prove that $\sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \ell(\boldsymbol{\beta}) \overset{a.s.}{\to} 0$ as $p \to \infty$. Define $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_j)_{j=1}^p \in \mathbb{R}^p$ as

$$\tilde{\beta}_j = p^{-2/3}(2y_{\lceil nj/p \rceil} - 1)x_{\lceil nj/p \rceil, j}, \qquad j = 1, \dots, p,$$

where $\lceil a \rceil$ denotes the smallest integer larger or equal to a. It follows that

$$p^{-1/3}\mathbf{x}_i^\intercal \tilde{\boldsymbol{\beta}} = p^{-1}(2y_i - 1)\sum\nolimits_{j\in D_i} x_{ij}^2 + p^{-1}\sum\nolimits_{j\notin D_i} \zeta_{ij},$$

where $D_i = \{j \in \{1, \dots, p\} : (i-1)p/n < j \leq ip/n\}$ and we defined $\zeta_{ij} = x_{ij}x_{\lceil nj/p \rceil, j}(2y_{\lceil nj/p \rceil} - 1)$. Since $(x_{ij}^2)_{j\in D_i}$ and $(\zeta_{ij})_{j\notin D_i}$ are independent variables with bounded variance, the size of D_i is asymptotic to $n^{-1}p$ as $p \to \infty$ and $\mathbb{E}(\zeta_{ij}) = 0$ for $j \notin D_i$, Lemma 2.9 implies that $\lim_{p\to\infty} p^{-1/3}\mathbf{x}_i^\intercal \tilde{\boldsymbol{\beta}} \overset{a.s.}{=} n^{-1}(2y_i - 1)\sigma_x^2$. Assuming $\sigma_x^2 > 0$ without loss of generality (when $\sigma_x^2 = 0$ it holds $\bar{\boldsymbol{\beta}}^* \overset{a.s.}{=} 0$) it follows that $\mathbf{x}_i^\intercal \tilde{\boldsymbol{\beta}} \overset{a.s.}{\to} +\infty$ if $y_i = 1$ and $\mathbf{x}_i^\intercal \tilde{\boldsymbol{\beta}} \overset{a.s.}{\to} -\infty$ if $y_i = 0$ as $p \to \infty$ and therefore $\sum_{i=1}^n \log \Phi[(2y_i - 1)\mathbf{x}_i^\intercal \tilde{\boldsymbol{\beta}}] \overset{a.s.}{\to} 0$ as $p \to \infty$. Moreover $\|\tilde{\boldsymbol{\beta}}\|^2 = p^{-1/3}(p^{-1}\sum_{j=1}^p x_{\lceil nj/p \rceil, j}^2) \overset{a.s.}{\to} 0$ as $p \to \infty$ by Lemma 2.9. Thus $0 \geq \sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \ell(\boldsymbol{\beta}) \geq \ell(\tilde{\boldsymbol{\beta}}) \overset{a.s.}{\to} 0$ as $p \to \infty$ as desired. □

Lemma 2.14. *Let q_1 and q_2 be probability distributions on $\mathbb{R}^p$. Then, for any $\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p$, we have* $\text{KL}[q_1 \,||\, q_2] \geq 2\left|\text{pr}_{q_1} - \text{pr}_{q_2}\right|^2$, *where* $\text{pr}_{q_l} = \int \Phi(\mathbf{x}_{\text{NEW}}^\intercal \boldsymbol{\beta}) q_l(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta}$ *for $l = 1, 2$.*

Proof. By Pinsker's inequality, $\text{KL}[q_1 \,||\, q_2] \geq 2\,\text{TV}[q_1, q_2]^2$ where $\text{TV}[\cdot, \cdot]$ denotes the total variation distance between probability distributions. Recall that it holds $\text{TV}[q_1, q_2] = \sup_{h:\mathbb{R}^p\to[0,1]} |\int_{\mathbb{R}^p} h(\boldsymbol{\beta}) q_1(\boldsymbol{\beta}) d\boldsymbol{\beta} - \int_{\mathbb{R}^p} h(\boldsymbol{\beta}) q_2(\boldsymbol{\beta}) d\boldsymbol{\beta}|$. Taking $h(\boldsymbol{\beta}) = \Phi(\mathbf{x}_{\text{NEW}}^\intercal \boldsymbol{\beta})$ in the above equation we obtain the desired statement. □

Proof of Theorem 2.1. As noted in Armagan and Zaretzki (2011), the CAVI algorithm for MF-VB is equivalent to an EM algorithm for $p(\boldsymbol{\beta}|\mathbf{y})$ with missing data $\mathbf{z}$, which in this case is guaranteed to converge to the unique maximizer of $p(\boldsymbol{\beta}|\mathbf{y})$ by, e.g., Theorem 3.2 of McLachlan and Krishnan (2007) and the fact that $p(\boldsymbol{\beta}|\mathbf{y})$ is strictly concave (Haberman, 1974). Therefore $\mathbb{E}_{q_{\text{MF}}^*(\boldsymbol{\beta})}(\boldsymbol{\beta}) = \bar{\boldsymbol{\beta}}^*$ and Lemma 2.13 implies that $\nu_p^{-1}\|\mathbb{E}_{q_{\text{MF}}^*(\boldsymbol{\beta})}(\boldsymbol{\beta})\| \overset{a.s.}{\to} 0$ as $p \to \infty$.

We now show that $\nu_p^{-2}\|\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{y})}(\boldsymbol{\beta})\|^2 \overset{a.s.}{\to} \frac{(\alpha\sigma_x^2)}{(1+\alpha\sigma_x^2)}c^2 n$ as $p \to \infty$. By the law of total expectation $\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{y})}(\boldsymbol{\beta}) = \mathbf{V}\mathbf{X}^\intercal \mathbb{E}_{p(\mathbf{z}|\mathbf{y})}(\mathbf{z})$. It follows that we have $\|\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{y})}(\boldsymbol{\beta})\|^2 = \mathbb{E}_{p(\mathbf{z}|\mathbf{y})}(\mathbf{z})^\intercal \mathbf{X}\mathbf{V}^\intercal \mathbf{V}\mathbf{X}^\intercal \mathbb{E}_{p(\mathbf{z}|\mathbf{y})}(\mathbf{z})$. Applying the Woodbury's identity to $\mathbf{V}$ we have $\mathbf{V}\mathbf{X}^\intercal = \nu_p^2 \mathbf{X}^\intercal(\mathbf{I}_n + \nu_p^2 \mathbf{X}\mathbf{X}^\intercal)^{-1}$. Thus, we can write $\mathbf{X}\mathbf{V}^\intercal\mathbf{V}\mathbf{X}^\intercal = (1+\sigma_x^2 p\nu_p^2)^{-2}\sigma_x^2 p\nu_p^4 \mathbf{S}^\intercal(\sigma_x^2 p)^{-1}\mathbf{X}\mathbf{X}^\intercal\mathbf{S}$ with $\mathbf{S} = (1 + \sigma_x^2 p\nu_p^2)(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal)^{-1}$. Since $\mathbf{S}^\intercal(\sigma_x^2 p)^{-1}\mathbf{X}\mathbf{X}^\intercal\mathbf{S} \overset{a.s.}{\to} \mathbf{I}_n$ as $p \to \infty$ from Lemma 2.10. Multiplying and dividing by the appropriate terms in the expression for $\|\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{y})}(\boldsymbol{\beta})\|^2$, it also follows that $\lim_{p\to\infty} \nu_p^{-2}\|\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{y})}(\boldsymbol{\beta})\|^2 \overset{a.s.}{=} \lim_{p\to\infty} \frac{(\sigma_x^2 p\nu_p^2)}{(1+\sigma_x^2 p\nu_p^2)}\|\mathbb{E}_{p(\mathbf{z}|\mathbf{y})}[(1+$

$\sigma_x^2 p\nu_p^2)^{-1/2}\mathbf{z}]\|^2$. Since it holds $(\mathbf{z} \mid \mathbf{y}) \sim \text{TN}[\mathbf{0}, (\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal), \mathbb{A}]$, we obtain $[(1+\sigma_x^2 p\nu_p^2)^{-1/2}\mathbf{z} \mid \mathbf{y}] \sim \text{TN}[\mathbf{0}, (1+\sigma_x^2 p\nu_p^2)^{-1}(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal), \mathbb{A}]$. Then,

$$\mathbb{E}_{p(z_i|\mathbf{y})}[(1+\sigma_x^2 p\nu_p^2)^{-1/2} z_i] = \frac{1}{\tilde{\psi}^{(p)}} \int_{\mathbb{A}} \tilde{u}_i \phi_n[\tilde{\mathbf{u}}; (1+\sigma_x^2 p\nu_p^2)^{-1}(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal)] d\tilde{\mathbf{u}},$$

where $\tilde{\psi}^{(p)} = \text{pr}(\mathbf{u}^{(p)} \in \mathbb{A})$ for $\mathbf{u}^{(p)} \sim \text{N}_n[\mathbf{0}, (1+\sigma_x^2 p\nu_p^2)^{-1}(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal)]$. Thus, Lemma 2.10 together with a domination argument similar to the one used in the proof of Lemma 2.12 imply that, as $p \to \infty$,

$$\mathbb{E}_{p(z_i|\mathbf{y})}[(1+\sigma_x^2 p\nu_p^2)^{-1/2} z_i] \overset{a.s.}{\to} 2^n \int_{\mathbb{A}} \tilde{u}_i \phi_n(\tilde{\mathbf{u}}; \mathbf{I}_n) d\tilde{\mathbf{u}} = c(2y_i - 1),$$

where $c = 2\int_0^\infty u\phi(u)du$. Therefore, $\lim_{p\to\infty} \nu_p^{-2} \|\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{y})}(\boldsymbol{\beta})\|^2 \overset{a.s.}{=} \frac{(\alpha\sigma_x^2)}{(1+\alpha\sigma_x^2)} \sum_{i=1}^n c^2 = \frac{(\alpha\sigma_x^2)}{(1+\alpha\sigma_x^2)} c^2 n$.

Finally, we show that $\liminf_{p\to\infty} \text{KL}[q^*_{\text{MF}}(\boldsymbol{\beta}) \mid\mid p(\boldsymbol{\beta} \mid \mathbf{y})] \overset{a.s.}{>} 0$. Lemma 2.14 implies $\text{KL}[q^*_{\text{MF}}(\boldsymbol{\beta}) \mid\mid p(\boldsymbol{\beta} \mid \mathbf{y})] \geq 2\,|\text{pr}_{\text{MF}} - \text{pr}_{\text{SUN}}|^2$, where $\text{pr}_{\text{SUN}} = \int \Phi(\mathbf{x}_{\text{NEW}}^\intercal\boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{y}) \text{d}\boldsymbol{\beta}$ and $\text{pr}_{\text{MF}} = \int \Phi(\mathbf{x}_{\text{NEW}}^\intercal\boldsymbol{\beta}) q^*_{\text{MF}}(\boldsymbol{\beta}) \text{d}\boldsymbol{\beta}$. To accomplish this goal, we consider $\mathbf{x}_{\text{NEW}} = (1+\sigma_x^2 p\nu_p^2)^{-1/2}\mathbf{X}^\intercal\mathbf{H}^{-1}\boldsymbol{\delta}$, with $\boldsymbol{\delta} = (2y_1 - 1, 0, \ldots, 0)^\intercal$, and show that $\lim_{p\to\infty} |\text{pr}_{\text{MF}} - \text{pr}_{\text{SUN}}| > 0$. Here we can assume without loss of generality that $\mathbf{H}$ is invertible because $\mathbf{H} \overset{a.s.}{\to} \mathbf{I}_n$ as $p \to \infty$ by Lemma 2.11 and the set of $n \times n$ non-singular matrices is open. This implies that $\mathbf{H}$ is eventually invertible as $p \to \infty$ almost surely. By definition of $\mathbf{x}_{\text{NEW}}$ we have

$$\begin{aligned}\nu_p^2 \|\mathbf{x}_{\text{NEW}}\|^2 &= \nu_p^2 \mathbf{x}_{\text{NEW}}^\intercal \mathbf{x}_{\text{NEW}} \\ &= \frac{\sigma_x^2 p\nu_p^2}{1+\sigma_x^2 p\nu_p^2} \boldsymbol{\delta}^\intercal \mathbf{H}^{-1} (\sigma_x^2 p)^{-1} \mathbf{X}\mathbf{X}^\intercal \mathbf{H}^{-1}\boldsymbol{\delta} \overset{a.s.}{\to} \frac{1+\alpha\sigma_x^2}{\alpha\sigma_x^2} \quad \text{as } p \to \infty,\end{aligned}$$

because $\mathbf{H}^{-1} \overset{a.s.}{\to} \frac{1+\alpha\sigma_x^2}{\alpha\sigma_x^2}\mathbf{I}_n$ and $(\sigma_x^2 p)^{-1}\mathbf{X}\mathbf{X}^\intercal \overset{a.s.}{\to} \mathbf{I}_n$ as $p \to \infty$ by Lemmas 2.11 and 2.10, respectively, and $\|\boldsymbol{\delta}\| = 1$. By (2.7) we have $\text{pr}_{\text{MF}} = \Phi[\mathbf{x}_{\text{NEW}}^\intercal \bar{\boldsymbol{\beta}}^* (1 + \mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{x}_{\text{NEW}})^{-1/2}]$, and by combining Cauchy-Schwarz inequality and $\mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{x}_{\text{NEW}} \geq 0$ we have that

$$|\mathbf{x}_{\text{NEW}}^\intercal \bar{\boldsymbol{\beta}}^* (1 + \mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{x}_{\text{NEW}})^{-1/2}| \leq \|\mathbf{x}_{\text{NEW}}\| \|\bar{\boldsymbol{\beta}}^*\| \overset{a.s.}{\to} 0,$$

as $p \to \infty$, where the latter convergence follows from $\nu_p \|\mathbf{x}_{\text{NEW}}\| \overset{a.s.}{\to} [(1+\alpha\sigma_x^2)/\alpha\sigma_x^2]^{1/2} \in (0, \infty)$ and $\nu_p^{-1}\|\bar{\boldsymbol{\beta}}^*\| \overset{a.s.}{\to} 0$. Thus $\text{pr}_{\text{MF}} \overset{a.s.}{\to} 0.5$ as $p \to \infty$.

Consider now pr_{SUN}. With derivations analogous to those of equation (2.11), we can express pr_{SUN} as $\text{pr}_{\text{SUN}} = \mathbb{E}_{p(\mathbf{z}|\mathbf{y})}\{\Phi[\mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{X}^\intercal \mathbf{z}(1 + \mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{x}_{\text{NEW}})^{-1/2}]\}$. By definition of $\mathbf{x}_{\text{NEW}}$, we have that $\mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{x}_{\text{NEW}} = (1+\sigma_x^2 p\nu_p^2)^{-1}\boldsymbol{\delta}^\intercal\mathbf{H}^{-1}\boldsymbol{\delta}$ and, as $p\to\infty$, $\mathbf{H}^{-1} \overset{a.s.}{\to} \frac{1+\alpha\sigma_x^2}{\alpha\sigma_x^2}\mathbf{I}_n$ by Lemma 2.11 and $\|\boldsymbol{\delta}\| = 1$. Thus, $\mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{x}_{\text{NEW}} \overset{a.s.}{\to} 0$ if $p\nu_p \to \infty$, while $\mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{x}_{\text{NEW}} \overset{a.s.}{\to} (\alpha\sigma_x^2)^{-1}$ if $p\nu_p \to \alpha \in (0, \infty)$. Moreover, $\mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{X}^\intercal \mathbf{z} = \boldsymbol{\delta}^\intercal[(1+\sigma_x^2 p\nu_p^2)^{-1/2}\mathbf{z}]$ and $(1+\sigma_x^2 p\nu_p^2)^{-1/2}\mathbf{z} \to \text{TN}(\mathbf{0}, \mathbf{I}_n, \mathbb{A})$ in distribution as $p \to \infty$, almost surely. Combining these results with Slutsky's lemma and the fact that $\Phi(\cdot)$ is bounded and continuous, it follows

that $\mathbb{E}_{p(\mathbf{z}|\mathbf{y})}\{\Phi[\mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{X}^\intercal \mathbf{z}(1+\mathbf{x}_{\text{NEW}}^\intercal \mathbf{V}\mathbf{x}_{\text{NEW}})^{-1/2}]\} \xrightarrow{a.s.} \mathbb{E}_{p(\tilde{\mathbf{z}})}\{\Phi[\alpha\sigma_x^2(1+\alpha\sigma_x^2)^{-1}\boldsymbol{\delta}^\intercal\tilde{\mathbf{z}}]\}$ with $\tilde{\mathbf{z}} \sim \text{TN}(\mathbf{0}, \mathbf{I}_n, \mathbb{A})$. Thus $\text{pr}_{\text{SUN}} \xrightarrow{a.s.} \mathbb{E}_{p(\tilde{z}_1)}\{\Phi[(2y_1-1)\alpha\sigma_x^2(1+\alpha\sigma_x^2)^{-1}\tilde{z}_1]\} = \int_0^\infty \Phi[\alpha\sigma_x^2(1+\alpha\sigma_x^2)^{-1}z]2\phi(z)dz > 0.5$ as $p \to \infty$. It follows that $\liminf_{p\to\infty} \text{KL}[q^*_{\text{MF}}(\boldsymbol{\beta}) \mid\mid p(\boldsymbol{\beta} \mid \mathbf{y})] \geq 2\lim_{p\to\infty}|\text{pr}_{\text{MF}} - \text{pr}_{\text{SUN}}|^2 > 0$ almost surely as $p \to \infty$. □

2.A.2 Proof of Theorem 2.3, Corollary 2.4 and Proposition 2.5

Proof of Theorem 2.3. Leveraging the chain rule of the KL divergence we obtain that $\text{KL}[q_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z}) \mid\mid p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})] = \text{KL}[q_{\text{PFM}}(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{y})] + \mathbb{E}_{q_{\text{PFM}}(\mathbf{z})}\{\text{KL}[q_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) \mid\mid p(\boldsymbol{\beta} \mid \mathbf{z})]\}$, where $q_{\text{PFM}}(\boldsymbol{\beta}|\mathbf{z})$ appears only in the second summand. This quantity is always non-negative and coincides with zero, for every $q_{\text{PFM}}(\mathbf{z})$, if and only if $q^*_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) = p(\boldsymbol{\beta} \mid \mathbf{z}) = \phi_p(\boldsymbol{\beta} - \mathbf{V}\mathbf{X}^\intercal\mathbf{z}; \mathbf{V})$.

The expression for $q^*_{\text{PFM}}(\mathbf{z}) = \prod_{i=1}^n q^*_{\text{PFM}}(z_i)$ is instead a direct consequence of the closure under conditioning property of the multivariate truncated Gaussian (Horrace, 2005; Holmes and Held, 2006). In particular, adapting the results in Holmes and Held (2006), it easily follows that

$$p(z_i \mid \mathbf{z}_{-i}, \mathbf{y}) \propto \phi[z_i - (1-\mathbf{x}_i^\intercal\mathbf{V}\mathbf{x}_i)^{-1}\mathbf{x}_i^\intercal\mathbf{V}\mathbf{X}_{-i}^\intercal\mathbf{z}_{-i}; (1-\mathbf{x}_i^\intercal\mathbf{V}\mathbf{x}_i)^{-1}]\mathbb{1}[(2y_i-1)z_i > 0],$$

for $i = 1, \ldots, n$, where $\mathbf{X}_{-i}$ is the design matrix without row i. To obtain the expression for $q^*_{\text{PFM}}(z_i)$, $i = 1, \ldots, n$, note that, recalling e.g., Blei et al. (2017), the optimal solution for $q_{\text{PFM}}(\mathbf{z})$ which minimizes $\text{KL}[q_{\text{PFM}}(\mathbf{z}) \mid\mid p(\mathbf{z} \mid \mathbf{y})]$ within family of distributions that factorize over $z_1, \ldots, z_n$ can be expressed as $\prod_{i=1}^n q^*_{\text{PFM}}(z_i)$ with $q^*_{\text{PFM}}(z_i) \propto \exp[\mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z}_{-i})}(\log[p(z_i \mid \mathbf{z}_{-i}, \mathbf{y})])]$ for every $i = 1, \ldots, n$. Combining such a result with the above expression for $p(z_i \mid \mathbf{z}_{-i}, \mathbf{y})$ we have that $\exp[\mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z}_{-i})}(\log[p(z_i \mid \mathbf{z}_{-i}, \mathbf{y})])]$ is proportional to

$$\exp\left[-\frac{z_i^2 - 2z_i(1-\mathbf{x}_i^\intercal\mathbf{V}\mathbf{x}_i)^{-1}\mathbf{x}_i^\intercal\mathbf{V}\mathbf{X}_{-i}^\intercal\mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z}_{-i})}(\mathbf{z}_{-i})}{2(1-\mathbf{x}_i^\intercal\mathbf{V}\mathbf{x}_i)^{-1}}\right]\mathbb{1}[(2y_i-1)z_i > 0],$$

for $i = 1, \ldots, n$.. The above quantity coincides with the kernel of a Gaussian distribution having variance $\sigma_i^{*2} = (1-\mathbf{x}_i^\intercal\mathbf{V}\mathbf{x}_i)^{-1}$, expectation $\mu_i^* = \sigma_i^{*2}\mathbf{x}_i^\intercal\mathbf{V}\mathbf{X}_{-i}^\intercal\mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z}_{-i})}(\mathbf{z}_{-i})$ and truncation below zero if $y_i = 1$ or above zero if $y_i = 0$. Hence, each $q^*_{\text{PFM}}(z_i)$ is the density of a truncated normal with parameters specified in Theorem 2.3. The proof is concluded after noticing that the expression for $\bar{z}_i^* = \mathbb{E}_{q^*_{\text{PFM}}(z_i)}(z_i)$, $i = 1, \ldots, n$, in Theorem 2.3 follows directly from the mean of truncated normals. □

Proof of Corollary 2.4. From (2.8), we have that $q^*_{\text{PFM}}(\boldsymbol{\beta})$ coincides with the density of a random variable that has the same distribution of $\tilde{\mathbf{u}}^{(0)} + \mathbf{V}\mathbf{X}^\intercal\tilde{\mathbf{u}}^{(1)}$, where $\tilde{\mathbf{u}}^{(0)} \sim \text{N}_p(\mathbf{0}, \mathbf{V})$ and $\tilde{\mathbf{u}}^{(1)}$ is from an n-variate Gaussian with mean vector $\boldsymbol{\mu}^*$, diagonal covariance matrix $\boldsymbol{\sigma}^{*2}$ and generic ith component truncated either below or above zero depending of the sign of $(2y_i - 1)$, for $i = 1, \ldots, n$. Since $\tilde{\mathbf{u}}^{(1)}$ has independent components, by standard

properties of univariate truncated normal variables we obtain

$$\tilde{\mathbf{u}}^{(0)} + \mathbf{V}\mathbf{X}^\intercal\tilde{\mathbf{u}}^{(1)} \stackrel{d}{=} \mathbf{u}^{(0)} + \mathbf{V}\mathbf{X}^\intercal\bar{\mathbf{Y}}\boldsymbol{\sigma}^*\mathbf{u}^{(1)}, \text{ with } \bar{\mathbf{Y}} = \text{diag}(2y_1 - 1, \ldots, 2y_n - 1),$$

where $\mathbf{u}^{(0)} \sim \text{N}_p(\mathbf{V}\mathbf{X}^\intercal\boldsymbol{\mu}^*, \mathbf{V})$ and $\mathbf{u}^{(1)}$ is an n-variate Gaussian with mean vector $\mathbf{0}$, covariance matrix $\mathbf{I}_n$, and truncation below $-\bar{\mathbf{Y}}\boldsymbol{\sigma}^{*^{-1}}\boldsymbol{\mu}^*$. Calling $\boldsymbol{\xi} = \mathbf{V}\mathbf{X}^\intercal\boldsymbol{\mu}^*$, $\boldsymbol{\Omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega} = \mathbf{V} + \mathbf{V}\mathbf{X}^\intercal\boldsymbol{\sigma}^{*2}\mathbf{X}\mathbf{V}$, $\boldsymbol{\Delta} = \boldsymbol{\omega}^{-1}\mathbf{V}\mathbf{X}^\intercal\bar{\mathbf{Y}}\boldsymbol{\sigma}^*$, $\boldsymbol{\gamma} = \bar{\mathbf{Y}}\boldsymbol{\sigma}^{*^{-1}}\boldsymbol{\mu}^*$ and $\boldsymbol{\Gamma} = \mathbf{I}_n$, as in Corollary 2.4, we have that

$$\mathbf{u}^{(0)} + \mathbf{V}\mathbf{X}^\intercal\bar{\mathbf{Y}}\boldsymbol{\sigma}^*\mathbf{u}^{(1)} \stackrel{d}{=} \boldsymbol{\xi} + \boldsymbol{\omega}(\bar{\mathbf{u}}^{(0)} + \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\bar{\mathbf{u}}^{(1)}),$$

with $\bar{\mathbf{u}}^{(0)} \sim \text{N}_p(\mathbf{0}, \bar{\boldsymbol{\Omega}} - \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}^\intercal)$, and $\bar{\mathbf{u}}^{(1)}$ distributed as a n-variate Gaussian random variable with mean vector $\mathbf{0}$, covariance matrix $\boldsymbol{\Gamma}$, and truncation below $-\boldsymbol{\gamma}$. Recalling Arellano-Valle and Azzalini (2006) and Azzalini and Capitanio (2014) such a stochastic representation coincides with the one of the unified skew-normal random variable $\text{SUN}_{p,n}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$. □

Proof of Proposition 2.5. To prove Proposition 2.5, first notice that by the results in equation (2.8) and in Theorem 2.3, $\mathbf{z} = (z_1, \ldots, z_n)^\intercal$ denotes a vector whose entries have independent truncated normal approximating densities. Hence, $\mathbb{E}_{q^*_{\text{PFM}}(z_i)}(z_i) = \bar{z}_i^*$ and $\text{var}_{q^*_{\text{PFM}}(z_i)}(z_i) = \sigma_i^{*2}[1 - (2y_i - 1)\eta_i^*\mu_i^*/\sigma_i^* - \eta_i^{*2}]$ with $\eta_i = \phi(\mu_i^*/\sigma_i^*)\Phi[(2y_i - 1)\mu_i^*/\sigma_i^*]^{-1}$ for $i = 1, \ldots, n$. Using the parameters defined in Theorem 2.3, $\text{var}_{q^*_{\text{PFM}}(z_i)}(z_i)$ can be also re-written as $\text{var}_{q^*_{\text{PFM}}(z_i)}(z_i) = \sigma_i^{*2} - (\bar{z}_i^* - \mu_i^*)\bar{z}_i^*$. Therefore, $\mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}(\mathbf{z}) = \bar{\mathbf{z}}^*$ and $\text{var}_{q^*_{\text{PFM}}(\mathbf{z})}(\mathbf{z}) = \text{diag}[\sigma_1^{*2} - (\bar{z}_1^* - \mu_1^*)\bar{z}_1^*, \ldots, \sigma_n^{*2} - (\bar{z}_n^* - \mu_n^*)\bar{z}_n^*]$, where $\bar{z}_i^*$, μ_i^* and σ_i^*, $i = 1, \ldots, n$ are defined in Theorem 2.3 and Corollary 2.4. Combining these results with equation (2.8), and using the law of iterated expectations we have

$$\begin{aligned}
\mathbb{E}_{q^*_{\text{PFM}}(\boldsymbol{\beta})}(\boldsymbol{\beta}) &= \mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}[\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{z})}(\boldsymbol{\beta})] = \mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}(\mathbf{V}\mathbf{X}^\intercal\mathbf{z}) \\
&= \mathbf{V}\mathbf{X}^\intercal\mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}(\mathbf{z}) = \mathbf{V}\mathbf{X}^\intercal\bar{\mathbf{z}}^*, \\
\text{var}_{q^*_{\text{PFM}}(\boldsymbol{\beta})}(\boldsymbol{\beta}) &= \mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}[\text{var}_{p(\boldsymbol{\beta}|\mathbf{z})}(\boldsymbol{\beta})] + \text{var}_{q^*_{\text{PFM}}(\mathbf{z})}[\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{z})}(\boldsymbol{\beta})] \\
&= \mathbf{V} + \mathbf{V}\mathbf{X}^\intercal\text{var}_{q^*_{\text{PFM}}(\mathbf{z})}(\mathbf{z})\mathbf{X}\mathbf{V} \\
&= \mathbf{V} + \mathbf{V}\mathbf{X}^\intercal\text{diag}[\sigma_1^{*2} - (\bar{z}_1^* - \mu_1^*)\bar{z}_1^*, \ldots, \sigma_n^{*2} - (\bar{z}_n^* - \mu_n^*)\bar{z}_n^*]\mathbf{X}\mathbf{V},
\end{aligned}$$

thus proving equation (2.10).

To prove equation (2.11) it suffices to notice that $\text{pr}_{\text{PFM}}(y_{\text{NEW}} = 1 \mid \mathbf{y}) = \mathbb{E}_{q^*_{\text{PFM}}(\boldsymbol{\beta})}[\Phi(\mathbf{x}_{\text{NEW}}^\intercal\boldsymbol{\beta})]$. Hence, by applying again the law of iterated expectations we have

$$\begin{aligned}
\mathbb{E}_{q^*_{\text{PFM}}(\boldsymbol{\beta})}[\Phi(\mathbf{x}_{\text{NEW}}^\intercal\boldsymbol{\beta})] &= \mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}\{\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{z})}[\Phi(\mathbf{x}_{\text{NEW}}^\intercal\boldsymbol{\beta})]\} \\
&= \mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}\{\Phi[\mathbf{x}_{\text{NEW}}^\intercal\mathbf{V}\mathbf{X}^\intercal\mathbf{z}(1 + \mathbf{x}_{\text{NEW}}^\intercal\mathbf{V}\mathbf{x}_{\text{NEW}})^{-1/2}]\}.
\end{aligned}$$

The last equality follows from the fact that $p(\boldsymbol{\beta} \mid \mathbf{z})$ is a Gaussian density and hence $\mathbb{E}_{p(\boldsymbol{\beta}|\mathbf{z})}[\Phi(\mathbf{x}_{\text{NEW}}^\intercal\boldsymbol{\beta})]$ can be derived in closed-form; see e.g., Lemma 7.1 in Azzalini and Capitanio (2014). □

2.A.3 Proof of Theorem 2.6 and Corollary 2.7

Proof of Theorem 2.6. As a consequence of the discussion after the statement of Theorem 2.3, the density $q^*_{\text{PFM}}(\mathbf{z})$ minimizes the KL divergence to $p(\mathbf{z}|\mathbf{y})$ within the family of distributions that factorize over $z_1, \dots, z_n$. Thus $\text{KL}[q^*_{\text{PFM}}(\mathbf{z})||p(\mathbf{z}|\mathbf{y})] \leq \text{KL}[\text{TN}(\mathbf{0}, (1 + \sigma_x^2 p \nu_p^2)\mathbf{I}_n, \mathbb{A})||p(\mathbf{z}|\mathbf{y})]$. Since the KL divergence is invariant with respect to bijective transformations and $p(\mathbf{z}|\mathbf{y}) = \text{TN}(\mathbf{0}, \mathbf{I}_n + \nu_p^2 \mathbf{X}\mathbf{X}^\intercal, \mathbb{A})$, then rescaling each z_i by $(1 + \sigma_x^2 p \nu_p^2)^{-1/2}$ we have

$$\text{KL}[\text{TN}(\mathbf{0}, (1+\sigma_x^2 p \nu_p^2)\mathbf{I}_n, \mathbb{A})||p(\mathbf{z}|\mathbf{y})] = \text{KL}[\text{TN}(\mathbf{0}, \mathbf{I}_n, \mathbb{A})||\text{TN}(\mathbf{0}, (1+\sigma_x^2 p \nu_p^2)^{-1}(\mathbf{I}_n+\nu_p^2 \mathbf{X}\mathbf{X}^\intercal), \mathbb{A}].$$

Lemma 2.10 shows that $(1 + \sigma_x^2 p \nu_p^2)^{-1}(\mathbf{I}_n + \nu_p^2 \mathbf{X}\mathbf{X}^\intercal) \overset{a.s.}{\to} \mathbf{I}_n$ and thus Lemma 2.12 implies that $\text{KL}[\text{TN}(\mathbf{0}, \mathbf{I}_n, \mathbb{A})||\text{TN}(\mathbf{0}, (1 + \sigma_x^2 p \nu_p^2)^{-1}(\mathbf{I}_n + \nu_p^2 \mathbf{X}\mathbf{X}^\intercal), \mathbb{A})] \overset{a.s.}{\to} 0$ as $p \to \infty$. From this result it follows that $\lim_{p\to\infty} \text{KL}[q^*_{\text{PFM}}(\mathbf{z})||p(\mathbf{z}|\mathbf{y})] \overset{a.s.}{=} 0$ as desired. □

Corollary 2.7. Lemma 2.14 and Theorem 2.6 also imply that

$$\sup_{\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p} |\text{pr}_{\text{PFM}} - \text{pr}_{\text{SUN}}| \leq \{\text{KL}[q^*_{\text{PFM}}(\boldsymbol{\beta}) \,||\, p(\boldsymbol{\beta} \mid \mathbf{y})]/2\}^{1/2} \overset{a.s.}{\to} 0$$

as $p \to \infty$. Moreover, in the proof of Theorem 2.1 it has been shown that setting $\mathbf{x}_{\text{NEW}} = (1 + \sigma_x^2 p \nu_p^2)^{-1/2} \mathbf{X}^\intercal \mathbf{H}^{-1} \boldsymbol{\delta}$ for every p leads to

$$\liminf_{p\to\infty} |\text{pr}_{\text{MF}} - \text{pr}_{\text{SUN}}| > 0,$$

almost surely, from which it follows the second part of the corollary. □

2.A.4 Proof of Theorem 2.8

Lemma 2.15. *Let $y \in \{0; 1\}$ be a generic binary response and call $\bar{z}^* = \mu^* + (2y - 1)\sigma^* \phi(\mu^*/\sigma^*)\Phi[(2y - 1)\mu^*/\sigma^*]^{-1}$, with $\mu^* \in \mathbb{R}$ and $\sigma^* \geq 0$. Then we have $\sup_{\mu^*,\sigma^*}(|\mu^*| + \sigma^*)^{-1}|\bar{z}^*| < \infty$.*

Proof. By the triangle inequality

$$(|\mu^*| + \sigma^*)^{-1}|\bar{z}^*| \leq 1 + (|\mu^*| + \sigma^*)^{-1}\sigma^* \phi(|\mu^*|/\sigma^*)/\Phi(-|\mu^*|/\sigma^*).$$

If $|\mu^*| \leq \sigma^*$ then $|\bar{z}^*|/(|\mu^*| + \sigma^*) \leq 1 + 1 \times \phi(0)/\Phi(-1) < \infty$. If $|\mu^*| > \sigma^*$, setting $t = |\mu^*|/\sigma^*$ and using the bound $\Phi(-t) \geq (2\pi)^{-1/2} t(t^2 + 1)^{-1} \exp(-t^2/2)$, which holds for every $t > 0$, we have

$$\begin{aligned}(|\mu^*| + \sigma^*)^{-1}|\bar{z}^*| &\leq 1 + |\mu^*|^{-1}\sigma^* \phi(t)/\Phi(-t) \\ &\leq 1 + t^{-1} \exp(-t^2/2)[(t^2 + 1)^{-1} t \exp(-t^2/2)]^{-1} \\ &= 1 + t^{-2}(t^2 + 1) < 3\end{aligned}$$

where in the last inequality we used $t > 1$. Combining the above results it follows that $\sup_{\mu^*,\sigma^*}(|\mu^*| + \sigma^*)^{-1}|\bar{z}^*| < \infty$ as desired. □

Lemma 2.16. *Under A1 and A2, for every $i = 1, \dots, n$, we have $(1+\sigma_x^2 p\nu_p^2)^{-1/2}\mu_i^{(1)} \overset{a.s.}{\to} 0$ as $p \to \infty$, where $\mu_i^{(1)}$ is defined as in Algorithm 2.*

Proof. **Case a)** $\alpha = \infty$. We show $p^{-1/2}\mu_i^{(1)} \overset{a.s.}{\to} 0$ as $p \to \infty$, from which the desired result is immediate. Lemma 2.15 implies $\sup_{\sigma_i^*} |\bar{z}_i^{(0)}|/\sigma_i^* < \infty$ and, since σ_i^* is almost surely asymptotic to $p^{1/2}$ as $p \to \infty$ by Lemma 2.11, it follows $\sup_{p\geq 1} p^{-1/2}|\bar{z}_i^{(0)}| \overset{a.s.}{<} \infty$ for every $i = 1, \dots, n$. Note that we are implicitly assuming Algorithm 2 to have fixed initialization $\mu_i^{(0)} \in \mathbb{R}$, $i = 1, \dots, n$. We now prove that $\lim_{p\to\infty} p^{-1/2}\mu_i^{(1)} \overset{a.s.}{=} 0$ and $\sup_{p\geq 1} p^{-1/2}|\bar{z}_i^{(1)}| \overset{a.s.}{<} \infty$ for every $i = 1, \dots, n$ by induction on i. When $i = 1$, recalling the definition of $\mu_1^{(1)}$ in Algorithm 2, we have that $|p^{-1/2}\mu_1^{(1)}| = |\sigma_1^{*2}\sum_{i'=2}^n H_{1i'}p^{-1/2}\bar{z}_{i'}^{(0)}| \leq \sum_{i'=2}^n \sigma_1^{*2}|H_{1i'}|p^{-1/2}|\bar{z}_{i'}^{(0)}|$. Lemma 2.11 and the fact that σ_1^{*2} is almost surely asymptotic to p imply that $\sigma_1^{*2}|H_{1i'}| \overset{a.s.}{\to} 0$ for every $i' \geq 2$ as $p \to \infty$. Combining the latter with $\sup_{p\geq 1} p^{-1/2}|\bar{z}_{i'}^{(0)}| \overset{a.s.}{<} \infty$ we obtain $p^{-1/2}\mu_1^{(1)} \overset{a.s.}{\to} 0$ as $p \to \infty$. Combining the latter with Lemma 2.15, we obtain $\sup_{p\geq 1} p^{-1/2}|\bar{z}_1^{(1)}| \overset{a.s.}{<} \infty$. We thus proved the desired statement for $i = 1$.

When $i > 1$, by simple manipulations of the expressions in Algorithm 2, we can express $\mu_i^{(1)}/\sigma_i^*$ as

$$p^{-1/2}\mu_i^{(1)} = \sum\nolimits_{i'=1}^{i-1} \sigma_i^{*2} H_{ii'} p^{-1/2}\bar{z}_{i'}^{(1)} + \sum\nolimits_{i'=i+1}^{n} \sigma_i^{*2} H_{ii'} p^{-1/2}\bar{z}_{i'}^{(0)}.$$

Now, for $i' > i$ we have $|\sigma_i^{*2} H_{ii'} p^{-1/2}\bar{z}_{i'}^{(0)}| \overset{a.s.}{\to} 0$ by the same arguments of the $i = 1$ case above. For $i' < i$ we have $|\sigma_i^{*2} H_{ii'} p^{-1/2}\bar{z}_{i'}^{(1)}| \overset{a.s.}{\to} 0$ by Lemma 2.11, the fact that σ_i^{*2} is almost surely asymptotic to $\sigma_x^2\nu_p^2 p$ and $\sup_{p\geq 1} p^{-1/2}|\bar{z}_{i'}^{(1)}| < \infty$ for $i' < i$ by the inductive hypothesis. It follows that $\lim_{p\to\infty} p^{-1/2}\mu_i^{(1)} \overset{a.s.}{=} 0$ and thus, by Lemma 2.15, also that $\sup_{p\geq 1} p^{-1/2}|\bar{z}_i^{(1)}| < \infty$ *a.s.*. The thesis follows by induction.

Case b) $\alpha \in (0, \infty)$. In such a case the stronger result $\mu_i^{(1)} \overset{a.s.}{\to} 0$ as $p \to \infty$ holds. The proof follows the same steps of the previous case. First, Lemma 2.15 implies $\sup_{\sigma_i^*} |\bar{z}_i^{(0)}|/\sigma_i^* < \infty$ and, since σ_i^* is almost surely asymptotic to $(1 + \alpha\sigma_x^2)^{1/2}$ as $p \to \infty$ by Lemma 2.11, it follows $\sup_{p\geq 1} |\bar{z}_i^{(0)}| \overset{a.s.}{<} \infty$ for every $i = 1, \dots, n$. Then, adapting the proof of the previous case, one can show by induction that $\lim_{p\to\infty} \mu_i^{(1)} \overset{a.s.}{=} 0$ and $\sup_{p\geq 1} |\bar{z}_i^{(1)}| \overset{a.s.}{<} \infty$ for every $i = 1, \dots, n$, and the proof is concluded. □

Proof of Theorem 2.8. The chain rule for KL divergences and the fact that $q_{\text{PFM}}^{(1)}(\boldsymbol{\beta}|\mathbf{z}) = p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{z})$ imply that $\text{KL}[q_{\text{PFM}}^{(1)}(\boldsymbol{\beta})||p(\boldsymbol{\beta}|\mathbf{y})] \leq \text{KL}[q_{\text{PFM}}^{(1)}(\boldsymbol{\beta}, \mathbf{z})||p(\boldsymbol{\beta}, \mathbf{z}|\mathbf{y})] = \text{KL}[q_{\text{PFM}}^{(1)}(\mathbf{z})||p(\mathbf{z}|\mathbf{y})]$. Since $q_{\text{PFM}}^{(1)}(\mathbf{z}) = \text{TN}(\boldsymbol{\mu}^{(1)}, \boldsymbol{\sigma}^{*2}, \mathbb{A})$ and $p(\mathbf{z}|\mathbf{y}) = \text{TN}[\mathbf{0}, (\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal), \mathbb{A}]$, then, calling $k_p = (1 + \sigma_x^2 p\nu_p)$, if we rescale by $k_p^{-1/2}$ we get

$$\text{KL}[q_{\text{PFM}}^{(1)}(\mathbf{z})||p(\mathbf{z}|\mathbf{y})] = \text{KL}[\text{TN}(k_p^{-1/2}\boldsymbol{\mu}^{(1)}, k_p^{-1}\boldsymbol{\sigma}^{*2}, \mathbb{A})||\text{TN}[\mathbf{0}, k_p^{-1}(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal), \mathbb{A}]].$$

Lemma 2.16 implies that $k_p^{-1/2}\boldsymbol{\mu}^{(1)} \overset{a.s.}{\to} \mathbf{0}$, while Lemmas 2.10 and 2.11 imply that both $k_p^{-1}(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal)$ and $k_p^{-1}\boldsymbol{\sigma}^{*2}$ converge a.s. to $\mathbf{I}_n$ as $p \to \infty$. Therefore, we obtain

$\text{KL}[\text{TN}(k_p^{-1/2}\boldsymbol{\mu}^{(1)}, k_p^{-1}\boldsymbol{\sigma}^{*2}, \mathbb{A})||\text{TN}[\mathbf{0}, k_p^{-1}(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal), \mathbb{A}]] \xrightarrow{a.s.} 0$ by Lemma 2.12, implying $\text{KL}[q^{(1)}_{\text{PFM}}(\boldsymbol{\beta})||p(\boldsymbol{\beta}|\mathbf{y})] \xrightarrow{a.s.} 0$ as desired. □

2.B Appendix: Computational cost of PFM-VB

We now discuss the computational cost of PFM-VB, showing that the whole routine requires matrix pre-computations with $\mathcal{O}(pn \cdot \min\{p, n\})$ cost and iterations with $\mathcal{O}(n \cdot \min\{p, n\})$ cost.

Consider first Algorithm 2. When $p \geq n$, one can pre-compute $\mathbf{X}\mathbf{V}\mathbf{X}^\intercal$ at $\mathcal{O}(pn^2)$ cost by applying the Woodbury's identity to $\mathbf{V}$, and then perform each iteration at $\mathcal{O}(n^2)$ cost. Instead, when $p < n$, one can pre-compute $\mathbf{X}\mathbf{V}$ at $\mathcal{O}(p^2n)$ cost, and then perform each iteration at $\mathcal{O}(pn)$ cost noting that

$$\mu_i^{(t)} = \sigma_i^{*2}\sum_{j=1}^{p}(\mathbf{X}\mathbf{V})_{ij}\alpha_j^{(t,i)}, \quad \text{with } \alpha_j^{(t,i)} = \sum_{i'=1}^{i-1} x_{i'j}\bar{z}_{i'}^{(t)} + \sum_{i'=i+1}^{n} x_{i'j}\bar{z}_{i'}^{(t-1)},$$

where the vector $\boldsymbol{\alpha}^{(t,i)} = (\alpha_1^{(t,i)}, \dots, \alpha_p^{(t,i)})^\intercal$ can be computed at $\mathcal{O}(p)$ cost from $\boldsymbol{\alpha}^{(t,i-1)}$ exploiting the recursive equations $\alpha_j^{(t,i)} = \alpha_j^{(t,i-1)} - x_{ij}z_i^{(t-1)} + x_{i-1,j}z_{i-1}^{(t)}$. Therefore, computing $\mu_i^{(t)}$ for $i = 1, \dots, n$, which is the most expensive part of Algorithm 2, can be done in $\mathcal{O}(np)$ operations using $\mathbf{X}\mathbf{V}$ and $\boldsymbol{\alpha}^{(t,i)}$. With simple calculations one can check that also computing $\text{ELBO}[q^{(t)}_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z})]$ requires $\mathcal{O}(n \cdot \min\{p, n\})$ operations, as it involves quadratic forms of $n \times n$ matrices with rank at most $\min\{p, n\}$; see https://github.com/augustofasano/Probit-PFMVB for the full ELBO expression.

Given the output of Algorithm 2, the mean of $\boldsymbol{\beta}$ under PFM-VB can be computed at $\mathcal{O}(pn \cdot \min\{p, n\})$ cost noting that, by (2.10), $\mathbb{E}_{q^*_{\text{PFM}}(\boldsymbol{\beta})}(\boldsymbol{\beta}) = \mathbf{V}\mathbf{X}^\intercal\bar{\mathbf{z}}^*$ and that $\mathbf{V}\mathbf{X}^\intercal$ can be computed at $\mathcal{O}(pn \cdot \min\{p, n\})$ cost using either its definition, when $p \leq n$, or the equality $\mathbf{V}\mathbf{X}^\intercal = \nu_p^2\mathbf{X}^\intercal\left(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^\intercal\right)^{-1}$, when $p > n$. Given $\mathbf{V}\mathbf{X}^\intercal$, one can compute the covariance matrix of $\boldsymbol{\beta}$ under PFM-VB at $\mathcal{O}(p^2n)$ cost using (2.10), and applying Woodbury's identity to $\mathbf{V}$ when $p > n$. On the other hand, the marginal variances $\text{var}_{q^*_{\text{PFM}}(\beta_j)}(\beta_j)$, $j = 1, \dots, p$, can be obtained at $\mathcal{O}(pn \cdot \min\{p, n\})$ cost by first computing $\mathbf{V}\mathbf{X}^\intercal$, and then exploiting (2.10) along with $V_{jj} = \nu_p^2\left[1 - \sum_{i=1}^{n}(\mathbf{V}\mathbf{X}^\intercal)_{ji}x_{ij}\right]$, which follows from $\mathbf{V}(\mathbf{I}_p + \nu_p^2\mathbf{X}^\intercal\mathbf{X}) = \nu_p^2\mathbf{I}_p$.

Finally, the Monte Carlo estimates of the approximate predictive probabilities $\text{pr}_{\text{PFM}}(y_{\text{NEW}} = 1 \mid \mathbf{y})$ in (2.11) can be computed at $\mathcal{O}(pn \cdot \min\{p, n\} + nR)$ cost, where R denotes the number of Monte Carlo samples. Indeed, simulating i.i.d. realizations $\mathbf{z}^{(r)}$, $r = 1, \dots, R$, from $q^*_{\text{PFM}}(\mathbf{z})$ for has an $\mathcal{O}(nR)$ cost, while computing $\Phi[\mathbf{x}_{\text{NEW}}^\intercal\mathbf{V}\mathbf{X}^\intercal\mathbf{z}^{(r)}(1+\mathbf{x}_{\text{NEW}}^\intercal\mathbf{V}\mathbf{x}_{\text{NEW}})^{-1/2}]$ for $r = 1, \dots, R$ has $\mathcal{O}(pn \cdot \min\{p, n\} + nR)$ cost because, given $\mathbf{V}\mathbf{X}^\intercal$, the computation of $\mathbf{x}_{\text{NEW}}^\intercal\mathbf{V}\mathbf{X}^\intercal\mathbf{z}^{(r)}$ for $r = 1, \dots, R$ requires $\mathcal{O}(pn + nR)$ operations, while the computation of $\mathbf{x}_{\text{NEW}}^\intercal\mathbf{V}\mathbf{x}_{\text{NEW}}$ can be done in $\mathcal{O}(pn \cdot \min\{p, n\})$ operations using either

its definition, when $p \leq n$, or Woodbury's identity on $\mathbf{V}$, when $p > n$, leading to $\mathbf{x}_{\text{NEW}}^{\intercal}\mathbf{V}\mathbf{x}_{\text{NEW}} = \nu_p^2\|\mathbf{x}_{\text{NEW}}\|^2 - \nu_p^2(\mathbf{X}\mathbf{x}_{\text{NEW}})^{\intercal}(\mathbf{I}_n + \nu_p^2\mathbf{X}\mathbf{X}^{\intercal})^{-1}(\mathbf{X}\mathbf{x}_{\text{NEW}})$.

Chapter 3

A Closed-Form Filter for Binary Time Series

3.1 Introduction

Despite the availability of several alternative approaches for dynamic inference and prediction of binary time series (MacDonald and Zucchini, 1997), state-space models provide a source of continuous interest due to their flexibility in accommodating a variety of representations and dependence structures via an interpretable formulation (West and Harrison, 2006; Petris et al., 2009; Durbin and Koopman, 2012). Let $\mathbf{y}_t = (y_{1t}, \ldots, y_{mt})^\intercal \in \{0;1\}^m$ denote a vector of binary event data at time t and define with $\boldsymbol{\beta}_t = (\beta_{1t}, \ldots, \beta_{pt})^\intercal \in \mathbb{R}^p$ the corresponding vector of state variables. Adapting the notation in Petris et al. (2009) to our setting, we aim to provide closed-form expressions for the filtering, predictive and smoothing distributions in the multivariate dynamic probit model

$$p(\mathbf{y}_t \mid \boldsymbol{\beta}_t) = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t), \tag{3.1}$$

$$\boldsymbol{\beta}_t = \mathbf{G}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t),\ t = 1\ldots, n, \quad \boldsymbol{\beta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0), \tag{3.2}$$

with dependence structure as defined by the directed acyclic graph displayed in Figure 3.1. In (3.1), $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ is the cumulative distribution function of the $\mathrm{N}_m(\mathbf{0}, \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ evaluated at $\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t$, with $\mathbf{B}_t = \mathrm{diag}(2y_{1t} - 1, \ldots, 2y_{mt} - 1)$ denoting the $m \times m$ sign matrix associated with $\mathbf{y}_t$, which defines the multivariate probit likelihood in equation (3.1). Model (3.1)–(3.2) is a natural generalization of univariate probit models to multivariate settings, as we will clarify in equations (3.3)–(3.5). The quantities $\mathbf{F}_t, \mathbf{V}_t, \mathbf{G}_t, \mathbf{W}_t, \mathbf{a}_0$ and $\mathbf{P}_0$ define, instead, known matrices controlling the location, scale and dependence structure in the state-space model (3.1)–(3.2). Estimation and inference for these matrices is, itself, a relevant problem which can be addressed both from a frequentist and a Bayesian perspective. Yet our focus is on providing exact results for inference on state variables and prediction of future binary events under (3.1)–(3.2). Hence,

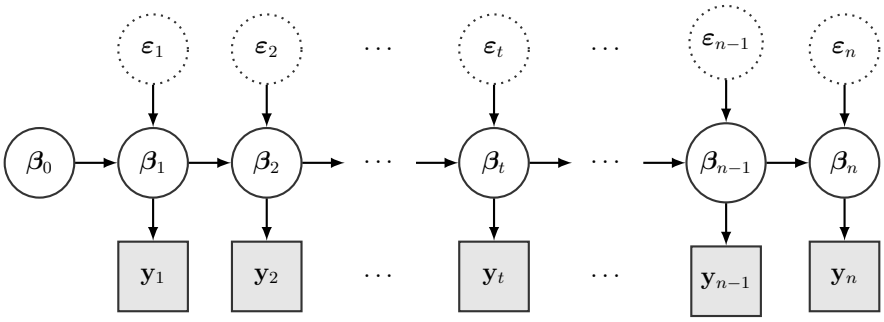

Figure 3.1: Representation of model (3.1)–(3.2). Dashed circles, solid circles and grey squares denote Gaussian errors, Gaussian states and observed binary data, respectively.

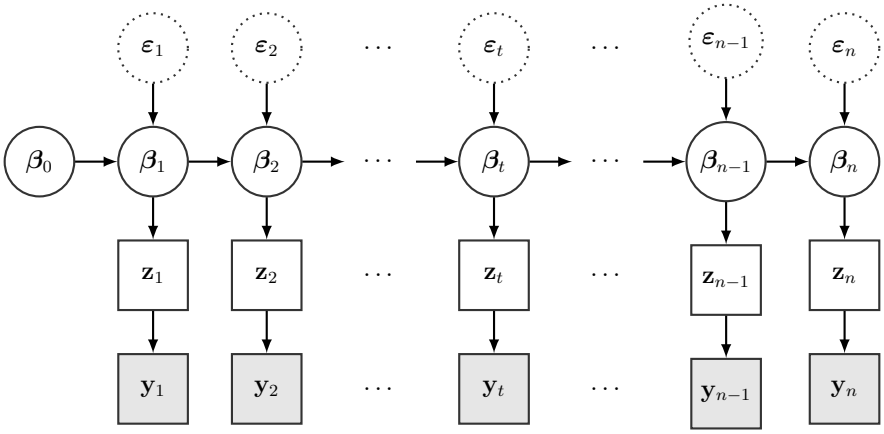

Figure 3.2: Representation of model (3.3)–(3.5). Dashed circles, solid circles, white squares and grey squares denote Gaussian errors, Gaussian states, latent Gaussian data and observed binary data, respectively.

consistent with the classical Kalman filter (Kalman, 1960), we rely on known system matrices $\mathbf{F}_t, \mathbf{V}_t, \mathbf{G}_t, \mathbf{W}_t, \mathbf{a}_0$ and $\mathbf{P}_0$. Nonetheless, results on marginal likelihoods, which can be used in parameter estimation, are provided in Section 3.3.2.

Model (3.1)–(3.2) provides a general representation encompassing a variety of formulations. For example, setting $\mathbf{V}_t = \mathbf{I}_m$ in (3.1) yields a standard probit regression, for $t = 1, \ldots, n$, which includes the classical univariate dynamic probit model when $m = 1$. These representations have appeared in several applications, especially within the econometrics literature, due to a direct connection between (3.1)–(3.2) and dynamic discrete choice models (Keane and Wolpin, 2009). This is due to the fact that representation (3.1)–(3.2) can be alternatively obtained via the dichotomization of an underlying state-space model for the m-variate Gaussian time series $\mathbf{z}_t = (z_{1t}, \ldots, z_{mt})^\intercal \in \mathbb{R}^m$, $t = 1, \ldots, n$, which is regarded, in econometric applications, as a set of time-varying utilities. Indeed, adapting the classical results from probit regression (Albert and Chib, 1993), model (3.1)–(3.2) is equivalent to

$$\mathbf{y}_t = (y_{1t}, \ldots, y_{mt})^\intercal = \mathbb{1}(\mathbf{z}_t > \mathbf{0}) = [\mathbb{1}(z_{1t} > 0), \ldots, \mathbb{1}(z_{mt} > 0)]^\intercal, \quad t = 1, \ldots, n, \quad (3.3)$$

with $\mathbf{z}_1, \ldots, \mathbf{z}_n$ evolving in time according to the Gaussian state-space model

$$p(\mathbf{z}_t \mid \boldsymbol{\beta}_t) = \phi_m(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{V}_t), \tag{3.4}$$

$$\boldsymbol{\beta}_t = \mathbf{G}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t),\ t = 1 \ldots, n, \quad \boldsymbol{\beta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0), \tag{3.5}$$

having dependence structure as defined by the directed acyclic graph displayed in Figure 3.2. In (3.4), $\phi_m(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{V}_t)$ denotes the density function of the Gaussian $\mathrm{N}_m(\mathbf{F}_t\boldsymbol{\beta}_t, \mathbf{V}_t)$ at point $\mathbf{z}_t \in \mathbb{R}^m$. To clarify the connection between (3.1)–(3.2) and model (3.3)–(3.5), note that the generic element $y_{lt} = \mathbb{1}(z_{lt} > 0)$ of $\mathbf{y}_t$ is 1 or 0 depending on whether $z_{lt} > 0$ or $z_{lt} \leq 0$. Therefore, $p(\mathbf{y}_t \mid \boldsymbol{\beta}_t) = \mathrm{pr}(\mathbf{B}_t\mathbf{z}_t > 0) = \mathrm{pr}[-\mathbf{B}_t(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\beta}_t) \leq \mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t] = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$, provided that $-\mathbf{B}_t(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\beta}_t) \sim \mathrm{N}_m(\mathbf{0}, \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ under (3.4).

As is clear from model (3.4)–(3.5), if $\mathbf{z}_{1:t} = (\mathbf{z}_1, \ldots, \mathbf{z}_t)$ were observed, dynamic inference on the states $\boldsymbol{\beta}_t$, for $t = 1, \ldots, n$, would be possible via direct application of the Kalman filter (Kalman, 1960). Indeed, exploiting Gaussian-Gaussian conjugacy and the conditional independence properties displayed in Figure 3.2, filtering $p(\boldsymbol{\beta}_t \mid \mathbf{z}_{1:t})$ and predictive $p(\boldsymbol{\beta}_t \mid \mathbf{z}_{1:t-1})$ distributions are also Gaussian and have parameters which can be computed recursively via simple expressions relying on the previous updates. Moreover, also the smoothing distribution $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{z}_{1:n})$ and its marginals $p(\boldsymbol{\beta}_t \mid \mathbf{z}_{1:n})$, $t \leq n$, can be obtained in closed-form leveraging the Gaussian-Gaussian conjugacy. However, in (3.3)–(3.5) only a dichotomized version $\mathbf{y}_t$ of $\mathbf{z}_t$ is available. Therefore the filtering, predictive and smoothing distributions of interest are $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n})$, respectively. Recalling model (3.1)–(3.2) and Bayes rule, the calculation of these quantities proceeds by updating the Gaussian distribution for the states in (3.2) with the probit likelihood in (3.1), thereby providing conditional distributions which seem not available in closed-form (Albert and Chib, 1993).

When the focus is online inference for filtering and prediction, a common solution to the above issue is to rely on approximations of model (3.1)–(3.2) which allow the implementation of standard Kalman filter updates, thus leading to approximate dynamic inference on the state variables via extended (Uhlmann, 1992) or unscented (Julier and Uhlmann, 1997) Kalman filters, among others. However, in different studies these approximations may lead to unreliable inference (Andrieu and Doucet, 2002). Markov chain Monte Carlo (MCMC) strategies (Carlin et al., 1992; Shephard, 1994; Soyer and Sung, 2013) address this problem, but, unlike the classical Kalman filter updates, these methods are suitable for batch learning of the smoothing distribution. Moreover, as discussed by Johndrow et al. (2019), common MCMC strategies face mixing or time-inefficiency issues, especially for imbalanced binary datasets. Sequential Monte Carlo solutions (Doucet et al., 2001) partially address MCMC issues and are specifically developed for online inference via particle-based representations of the conditional states distributions, which are propagated in time for dynamic filtering and prediction (Gordon et al., 1993; Kitagawa, 1996; Liu and Chen, 1998; Pitt and Shephard, 1999; Doucet et al., 2000; Andrieu and Doucet, 2002). These

methods provide state-of-the-art solutions in non-Gaussian state-space models, and can be also adapted to provide batch learning of the smoothing distribution; see Doucet and Johansen (2009) for a discussion on particle degeneracy issues that may arise in this setting. Nonetheless, sequential Monte Carlo is clearly still sub-optimal compared to the case in which $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n})$ are available in closed-form and belong to a tractable class of distributions whose parameters can be sequentially updated in time via simple analytical expressions.

In Section 3.3, we prove that for the dynamic probit model defined in (3.1)–(3.2) the quantities $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n})$ are unified skew-normal (SUN) distributions (Arellano-Valle and Azzalini, 2006) having tractable expressions for the recursive computation of the corresponding parameters. To the best of our knowledge, this result provides the first closed-form filter and smoother for binary time series, and allows improvements both in online and in batch inference within this framework. As highlighted in Section 3.2, the multivariate SUN distribution has several closure properties (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014) in addition to explicit formulas—involving the cumulative distribution function of multivariate normals—for the moments (Azzalini and Bacchieri, 2010; Gupta et al., 2013) and the normalizing constant (Arellano-Valle and Azzalini, 2006). In Sections 3.3, we exploit these properties to derive closed-form expressions for key functionals of $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n})$, including, in particular, the observations' predictive distribution $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ and the marginal likelihood $p(\mathbf{y}_{1:n})$. Besides these analytical results, we further propose in Section 3.4.1 an exact Monte Carlo scheme to compute complex functionals of the smoothing distribution. This routine relies on a stochastic representation of the SUN via a linear combination of Gaussians and truncated Gaussians (Arellano-Valle and Azzalini, 2006), and can be also applied effectively to calculate complex functionals of filtering and predictive distributions when the dimension of the time series is small-to-moderate, a common situation in several studies. As discussed in Section 3.4.2, the aforementioned strategies face computational bottlenecks in higher dimensional settings (Botev, 2017), due to challenges in computing cumulative distribution functions of multivariate Gaussians and in sampling from multivariate truncated normals. In these contexts, we propose a novel particle filter which exploits the SUN properties to obtain an optimal (Doucet et al., 2000) sequential Monte Carlo which effectively scales with t; see Section 3.4.2. As outlined in an illustrative study in Section 3.5, the methods developed in this chapter improve current strategies for batch and online inference in dynamic probit models. Future directions of research are discussed in Section 3.6.

3.2 The Unified Skew-Normal Distribution

Before deriving the filtering, predictive and smoothing distributions of model (3.1)–(3.2), let us first briefly review the SUN random variable. Arellano-Valle and Azzalini (2006) proposed this class to unify different generalizations (Arnold and Beaver, 2000; Arnold et al., 2002; Gupta et al., 2004; González-Farías et al., 2004) of the original multivariate skew-normal (Azzalini and Dalla Valle, 1996), whose density is obtained as the product of a multivariate Gaussian density and the cumulative distribution function of a standard normal evaluated at a value which depends on a skewness inducing vector of parameters. Motivated by the success of this formulation and of its various generalizations (Azzalini and Capitanio, 1999), Arellano-Valle and Azzalini (2006) developed a unifying representation, namely the unified skew-normal distribution. A random vector $\boldsymbol{\beta} \in \mathbb{R}^p$ has a unified skew-normal distribution, $\boldsymbol{\beta} \sim \text{SUN}_{p,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, if its density function can be expressed as

$$p(\boldsymbol{\beta}) = \phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \frac{\Phi_h[\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta}]}{\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})}, \tag{3.6}$$

where the covariance matrix of the Gaussian density $\phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega})$ is obtained as $\boldsymbol{\Omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}$, that is by re-scaling a correlation matrix $\bar{\boldsymbol{\Omega}}$ via a positive diagonal scale matrix $\boldsymbol{\omega} = (\boldsymbol{\Omega} \circ \mathbf{I}_p)^{1/2}$, with $\circ$ denoting the element-wise Hadamard product. Observe that the quantities p and h are not parameters, but define the dimensions of the multivariate Gaussian density and cumulative distribution function appearing in (3.6), respectively. Moreover, the dimensionality of the former coincides with that of the vector $\boldsymbol{\theta}$. In (3.6), the skewness inducing mechanism is driven by the cumulative distribution function of the $\text{N}_h(\mathbf{0}, \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta})$ computed at $\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi})$, whereas $\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})$ denotes the normalizing constant obtained by evaluating the cumulative distribution function of a $\text{N}_h(\mathbf{0}, \boldsymbol{\Gamma})$ at $\boldsymbol{\gamma}$. Arellano-Valle and Azzalini (2006) added a further identifiability condition which restricts the matrix $\boldsymbol{\Omega}^*$, with blocks $\boldsymbol{\Omega}^*_{[11]} = \boldsymbol{\Gamma}$, $\boldsymbol{\Omega}^*_{[22]} = \bar{\boldsymbol{\Omega}}$ and $\boldsymbol{\Omega}^*_{[21]} = \boldsymbol{\Omega}^{*\intercal}_{[12]} = \boldsymbol{\Delta}$, to be a full–rank correlation matrix.

To clarify the role of the parameters in expression (3.6), let us discuss a generative representation of the SUN. In particular, if $\mathbf{z}_0 \in \mathbb{R}^h$ and $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ characterize two random vectors jointly distributed as a $\text{N}_{h+p}(\mathbf{0}, \boldsymbol{\Omega}^*)$, then $\boldsymbol{\xi} + \boldsymbol{\omega}(\boldsymbol{\beta}_0 \mid \mathbf{z}_0 + \boldsymbol{\gamma} > \mathbf{0}) \sim \text{SUN}_{p,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ (Arellano-Valle and Azzalini, 2006). Hence, $\boldsymbol{\xi}$ and $\boldsymbol{\omega}$ control location and scale, respectively, whereas $\boldsymbol{\Gamma}$, $\bar{\boldsymbol{\Omega}}$ and $\boldsymbol{\Delta}$ define the dependence within $\mathbf{z}_0 \in \mathbb{R}^h$, within $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and between these two random vectors, respectively. Finally, $\boldsymbol{\gamma}$ controls the truncation in the partially observed Gaussian vector $\mathbf{z}_0 \in \mathbb{R}^h$. The above representation provides also key insights on our closed-form filter for the dynamic probit model (3.1)–(3.2). Indeed, according to (3.3)–(3.5), the filtering, predictive and smoothing distributions induced by model (3.1)–(3.2) can be also defined as $p[\boldsymbol{\beta}_t \mid \mathbb{1}(\mathbf{z}_1 > \mathbf{0}), \ldots, \mathbb{1}(\mathbf{z}_t > \mathbf{0})]$, $p[\boldsymbol{\beta}_t \mid \mathbb{1}(\mathbf{z}_1 > \mathbf{0}), \ldots, \mathbb{1}(\mathbf{z}_{t-1} > \mathbf{0})]$ and $p[\boldsymbol{\beta}_{1:n} \mid \mathbb{1}(\mathbf{z}_1 > \mathbf{0}), \ldots, \mathbb{1}(\mathbf{z}_n > \mathbf{0})]$, respec-

tively, with $(\mathbf{z}_t, \boldsymbol{\beta}_t)$ from the Gaussian state-space model (3.4)–(3.5) for $t = 1, \ldots, n$, thus highlighting the direct connection between these distributions and the generative representation of the SUN.

Another fundamental stochastic representation of the SUN distribution relies on linear combinations of Gaussian and truncated Gaussian random variables, thereby facilitating sampling from the SUN. In particular, recalling Azzalini and Capitanio (2014, Chapter 7.1.2) and Arellano-Valle and Azzalini (2006), if $\boldsymbol{\beta} \sim \text{SUN}_{p,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, then

$$\boldsymbol{\beta} \stackrel{d}{=} \boldsymbol{\xi} + \boldsymbol{\omega}(\mathbf{U}_0 + \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\mathbf{U}_1), \quad \mathbf{U}_0 \perp \mathbf{U}_1, \tag{3.7}$$

with $\mathbf{U}_0 \sim \text{N}_p(\mathbf{0}, \bar{\boldsymbol{\Omega}} - \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}^\intercal)$ and $\mathbf{U}_1$ from a $\text{N}_h(\mathbf{0}, \boldsymbol{\Gamma})$ truncated below $-\boldsymbol{\gamma}$. As we will clarify in Section 3.4, this result can facilitate efficient Monte Carlo inference on complex functionals of filtering, predictive and smoothing distributions in model (3.1)–(3.2), based on sampling from the corresponding SUN variable. Indeed, although key moments can be explicitly derived via direct differentiation of the SUN moment generating function (Arellano-Valle and Azzalini, 2006; Gupta et al., 2013), such a strategy requires tedious calculations in the unified skew-normal framework, when the focus is on complex functionals. Moreover, recalling Azzalini and Bacchieri (2010) and Gupta et al. (2013), the first and second order moments further require the evaluation of h-variate Gaussian cumulative distribution functions $\Phi_h(\cdot)$, thus affecting computational tractability in large h settings (Botev, 2017). In these situations, Monte Carlo integration provides an effective solution, especially when independent samples can be generated efficiently. Therefore, we mostly focus on improved Monte Carlo inference in model (3.1)–(3.2) exploiting the SUN properties, and refer to Azzalini and Bacchieri (2010) and Gupta et al. (2013) for a closed-form expression of the expectation, variance and cumulative distribution function of SUN variables. As clarified in Section 3.3, h increases linearly with time t in the SUN filtering and predictive distributions.

Before concluding our overview, we shall emphasize that unified skew-normal random variables are also closed under marginalization, linear combinations and conditioning (Azzalini and Capitanio, 2014). These properties facilitate the derivation of the SUN filtering, predictive and smoothing distributions in model (3.1)–(3.2).

3.3 Filtering, Prediction and Smoothing

In this section, it is shown that all the distributions of interest admit a closed-form SUN representation. In particular, in Section 3.3.1 we prove that closed-form filters—meant here as exact updating schemes for predictive and filtering distributions based on simple recursive expressions for the associated parameters—can be derived for model (3.1)–(3.2), whereas in Section 3.3.2 we present the form of the SUN smoothing distribution and some

important consequences. The associated computational methods are then discussed in Section 3.4.

3.3.1 Filtering and Predictive Distributions

To obtain the exact form of the filtering and predictive distributions under (3.1)–(3.2), let us start from $p(\boldsymbol{\beta}_1 \mid \mathbf{y}_1)$. This first quantity characterizes the initial step of the filter recursion, and its derivation in Lemma 3.1 provides key intuitions to obtain the state predictive $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})$ and filtering $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$, for every $t \geq 2$. Lemma 3.1 states that $p(\boldsymbol{\beta}_1 \mid \mathbf{y}_1)$ is a SUN distribution. In the following, consistent with the notation of Section 3.2, whenever $\boldsymbol{\Omega}$ is a $p \times p$ covariance matrix, the associated matrices $\boldsymbol{\omega}$ and $\bar{\boldsymbol{\Omega}}$ are defined as $\boldsymbol{\omega} = (\boldsymbol{\Omega} \circ \mathbf{I}_p)^{1/2}$ and $\bar{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}^{-1}$, respectively. All proofs can be found in the Appendix and consider conjugacy properties of the SUN in probit models. Early findings on this result have been explored by Durante (2019) in the context of static univariate Bayesian probit regression. Here, we take a substantially different perspective by focusing on online inference in both multivariate and time–varying probit models that require novel and non–straightforward extensions. As seen in Soyer and Sung (2013) and Chib and Greenberg (1998), the increased complexity of this endeavor typically motivates a separate treatment relative to the static univariate case.

Lemma 3.1. *Under the dynamic probit model* (3.1)–(3.2)*, the first-step filtering distribution is*

$$(\boldsymbol{\beta}_1 \mid \mathbf{y}_1) \sim \text{SUN}_{p,m}(\boldsymbol{\xi}_{1|1}, \boldsymbol{\Omega}_{1|1}, \boldsymbol{\Delta}_{1|1}, \boldsymbol{\gamma}_{1|1}, \boldsymbol{\Gamma}_{1|1}), \tag{3.8}$$

with parameters $\boldsymbol{\xi}_{1|1} = \mathbf{G}_1\mathbf{a}_0$, $\boldsymbol{\Omega}_{1|1} = \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1$, $\boldsymbol{\Delta}_{1|1} = \bar{\boldsymbol{\Omega}}_{1|1}\boldsymbol{\omega}_{1|1}\mathbf{F}_1^\intercal\mathbf{B}_1\mathbf{s}_1^{-1}$, $\boldsymbol{\gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1}$ *and* $\boldsymbol{\Gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1\mathbf{s}_1^{-1}$, *where* $\mathbf{s}_1 = [(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1) \circ \mathbf{I}_m]^{\frac{1}{2}}$.

Hence $p(\boldsymbol{\beta}_1 \mid \mathbf{y}_1)$ is a SUN distribution and its parameters can be obtained via tractable arithmetic expressions applied to the quantities characterizing model (3.1)–(3.2). Exploiting the results in Lemma 3.1, the general filter updates for the multivariate probit model can be obtained by induction for $t \geq 2$ and are presented in Theorem 3.2.

Theorem 3.2. *Let* $(\boldsymbol{\beta}_{t-1} \mid \mathbf{y}_{1:t-1}) \sim \text{SUN}_{p,m\cdot(t-1)}(\boldsymbol{\xi}_{t-1|t-1}, \boldsymbol{\Omega}_{t-1|t-1}, \boldsymbol{\Delta}_{t-1|t-1}, \boldsymbol{\gamma}_{t-1|t-1}, \boldsymbol{\Gamma}_{t-1|t-1})$ *be the filtering distribution at* $t-1$ *under* (3.1)–(3.2)*. Then the one-step-ahead state predictive distribution at* t *is*

$$(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1}) \sim \text{SUN}_{p,m\cdot(t-1)}(\boldsymbol{\xi}_{t|t-1}, \boldsymbol{\Omega}_{t|t-1}, \boldsymbol{\Delta}_{t|t-1}, \boldsymbol{\gamma}_{t|t-1}, \boldsymbol{\Gamma}_{t|t-1}), \tag{3.9}$$

with $\boldsymbol{\xi}_{t|t-1} = \mathbf{G}_t\boldsymbol{\xi}_{t-1|t-1}$, $\boldsymbol{\Omega}_{t|t-1} = \mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal + \mathbf{W}_t$, $\boldsymbol{\Delta}_{t|t-1} = \boldsymbol{\omega}_{t|t-1}^{-1}\mathbf{G}_t\boldsymbol{\omega}_{t-1|t-1}\boldsymbol{\Delta}_{t-1|t-1}$, $\boldsymbol{\gamma}_{t|t-1} = \boldsymbol{\gamma}_{t-1|t-1}$ *and* $\boldsymbol{\Gamma}_{t|t-1} = \boldsymbol{\Gamma}_{t-1|t-1}$*. Moreover, the filtering distribution at time* t *is*

$$(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t}) \sim \text{SUN}_{p,m\cdot t}(\boldsymbol{\xi}_{t|t}, \boldsymbol{\Omega}_{t|t}, \boldsymbol{\Delta}_{t|t}, \boldsymbol{\gamma}_{t|t}, \boldsymbol{\Gamma}_{t|t}), \tag{3.10}$$

with $\boldsymbol{\xi}_{t|t} = \boldsymbol{\xi}_{t|t-1}$, $\boldsymbol{\Omega}_{t|t} = \boldsymbol{\Omega}_{t|t-1}$, $\boldsymbol{\Delta}_{t|t} = [\boldsymbol{\Delta}_{t|t-1}, \bar{\boldsymbol{\Omega}}_{t|t}\boldsymbol{\omega}_{t|t}\mathbf{F}_t^\intercal\mathbf{B}_t\mathbf{s}_t^{-1}]$, $\boldsymbol{\gamma}_{t|t} = [\boldsymbol{\gamma}_{t|t-1}^\intercal, \boldsymbol{\xi}_{t|t}^\intercal\mathbf{F}_t^\intercal\mathbf{B}_t\mathbf{s}_t^{-1}]^\intercal$ *and* $\boldsymbol{\Gamma}_{t|t}$ *characterizes a full-rank correlation matrix with blocks* $\boldsymbol{\Gamma}_{t|t[11]} = \boldsymbol{\Gamma}_{t|t-1}$, $\boldsymbol{\Gamma}_{t|t[22]} = \mathbf{s}_t^{-1}\mathbf{B}_t(\mathbf{F}_t\boldsymbol{\Omega}_{t|t}\mathbf{F}_t^\intercal + \mathbf{V}_t)\mathbf{B}_t\mathbf{s}_t^{-1}$ *and* $\boldsymbol{\Gamma}_{t|t[21]} = \boldsymbol{\Gamma}_{t|t[12]}^\intercal = \mathbf{s}_t^{-1}\mathbf{B}_t\mathbf{F}_t\boldsymbol{\omega}_{t|t-1}\boldsymbol{\Delta}_{t|t-1}$, *where* $\mathbf{s}_t$ *is defined as* $\mathbf{s}_t = [(\mathbf{F}_t\boldsymbol{\Omega}_{t|t}\mathbf{F}_t^\intercal + \mathbf{V}_t) \circ \mathbf{I}_m]^{\frac{1}{2}}$.

Consistent with Theorem 3.2, online prediction and filtering in the multivariate dynamic probit model (3.1)–(3.2) proceeds by iterating between equations (3.9) and (3.10) as new observations stream in with time t. Both steps are based on closed-form distributions and rely on analytical expressions for recursive updating of the corresponding parameters as a function of the previous ones, thus providing an analog of the classical Kalman filter.

We also provide closed-form results for the predictive distribution of the multivariate binary data. Indeed, the prediction of future events $\mathbf{y}_t \in \{0;1\}^m$ given the current data $\mathbf{y}_{1:t-1}$, is a primary goal in applications of dynamic probit models. In our setting, this task requires the derivation of the predictive distribution $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ which coincides with $\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})\mathrm{d}\boldsymbol{\beta}_t$ in model (3.1)–(3.2), where $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})$ is the state predictive distribution in (3.9). Corollary 3.3 shows that this quantity has an explicit form.

Corollary 3.3. *Under model* (3.1)–(3.2)*, the observation predictive distribution* $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ *is*

$$p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}) = \frac{\Phi_{m\cdot t}(\boldsymbol{\gamma}_{t|t}; \boldsymbol{\Gamma}_{t|t})}{\Phi_{m\cdot(t-1)}(\boldsymbol{\gamma}_{t|t-1}; \boldsymbol{\Gamma}_{t|t-1})}, \tag{3.11}$$

for every time t*, with parameters* $\boldsymbol{\gamma}_{t|t}$, $\boldsymbol{\Gamma}_{t|t}$, $\boldsymbol{\gamma}_{t|t-1}$ *and* $\boldsymbol{\Gamma}_{t|t-1}$*, defined as in Theorem 3.2.*

Hence, the evaluation of probabilities of future events is possible via explicit calculations after marginalizing out analytically the predictive distribution of the states. As is clear from (3.11), this approach requires the calculation of Gaussian cumulative distribution functions whose dimension increases with t and m. Efficient evaluation of these integrals is possible for small-to-moderate t and m via recent minimax tilting (Botev, 2017). However, these methods are impractical when t and m are large. In Section 3.4, we develop new Monte Carlo methods based on independent samples and sequential Monte Carlo strategies to overcome this issue and allow scalable inference exploiting Theorem 3.2 to improve current solutions.

3.3.2 Smoothing Distribution

We now turn our focus to the smoothing distribution. In this case, the whole data $\mathbf{y}_{1:n}$ are available and the interest is on the distribution of either the whole sequence of the states $\boldsymbol{\beta}_{1:n}$ or a subset of it, given $\mathbf{y}_{1:n}$. Theorem 3.4 shows that also the smoothing

distribution $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n})$ belongs to the SUN family, and direct consequences of such a result, involving marginal smoothing and marginal likelihoods are reported in Corollaries 3.5 and 3.6, respectively.

Before stating the result, let us first introduce two block-diagonal matrices, $\mathbf{D}$ and $\mathbf{V}$, having dimensions $(m \cdot n) \times (p \cdot n)$ and $(m \cdot n) \times (m \cdot n)$ respectively, with diagonal blocks $\mathbf{D}_{[t,t]} = \mathbf{B}_t\mathbf{F}_t \in \mathbb{R}^{m \times p}$ and $\mathbf{V}_{[t,t]} = \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t \in \mathbb{R}^{m \times m}$, for every $t = 1, \ldots, n$. Moreover, let $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$ denote the mean and covariance matrix of the multivariate Gaussian for $\boldsymbol{\beta}_{1:n}$ induced by the state equations. Under (3.2), $\boldsymbol{\xi}$ is a $(p \cdot n) \times 1$ vector obtained by stacking the p-dimensional blocks $\boldsymbol{\xi}_{[t]} = \mathrm{E}(\boldsymbol{\beta}_t) = \mathbf{G}_1^t\mathbf{a}_0 \in \mathbb{R}^p$ for every $t = 1, \ldots, n$, with $\mathbf{G}_1^t = \mathbf{G}_t \cdots \mathbf{G}_1$. Similarly, letting $\mathbf{G}_q^t = \mathbf{G}_t \cdots \mathbf{G}_q$, also the $(p \cdot n) \times (p \cdot n)$ covariance matrix $\boldsymbol{\Omega}$ has a block structure with $(p \times p)$-dimensional blocks $\boldsymbol{\Omega}_{[t,t]} = \mathrm{var}(\boldsymbol{\beta}_t) = \mathbf{G}_1^t\mathbf{P}_0\mathbf{G}_1^{t\intercal} + \sum_{q=2}^{t} \mathbf{G}_q^t\mathbf{W}_{q-1}\mathbf{G}_q^{t\intercal} + \mathbf{W}_t$, for $t = 1, \ldots, n$, and $\boldsymbol{\Omega}_{[t,q]} = \boldsymbol{\Omega}_{[q,t]}^{\intercal} = \mathrm{cov}(\boldsymbol{\beta}_t, \boldsymbol{\beta}_q) = \mathbf{G}_{q+1}^t\boldsymbol{\Omega}_{[q,q]}$, for $t > q$.

Theorem 3.4. *Under model* (3.1)–(3.2)*, the joint smoothing distribution is*

$$(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n}) \sim \mathrm{SUN}_{p \cdot n, m \cdot n}(\boldsymbol{\xi}_{1:n|n}, \boldsymbol{\Omega}_{1:n|n}, \boldsymbol{\Delta}_{1:n|n}, \boldsymbol{\gamma}_{1:n|n}, \boldsymbol{\Gamma}_{1:n|n}), \tag{3.12}$$

where $\boldsymbol{\xi}_{1:n|n} = \boldsymbol{\xi}$, $\boldsymbol{\Omega}_{1:n|n} = \boldsymbol{\Omega}$, $\boldsymbol{\Delta}_{1:n|n} = \bar{\boldsymbol{\Omega}}\boldsymbol{\omega}\mathbf{D}^{\intercal}\mathbf{s}^{-1}$, $\boldsymbol{\gamma}_{1:n|n} = \mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}$, $\boldsymbol{\Gamma}_{1:n|n} = \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^{\intercal} + \mathbf{V})\mathbf{s}^{-1}$ *and* $\mathbf{s} = [(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^{\intercal} + \mathbf{V}) \circ \mathbf{I}_{m \cdot n}]^{1/2}$.

Since the SUN is closed under marginalization and linear combinations, it follows from Theorem 3.4 that the smoothing distribution for any combination of states is still a SUN. In particular, direct application of the results in Azzalini and Capitanio (2014, Chapter 7.1.2) provides the marginal smoothing distribution for any state $\boldsymbol{\beta}_t$ reported in Corollary 3.5.

Corollary 3.5. *Under model* (3.1)–(3.2)*, the marginal smoothing distribution at time* t *is*

$$(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:n}) \sim \mathrm{SUN}_{p, m \cdot n}(\boldsymbol{\xi}_{t|n}, \boldsymbol{\Omega}_{t|n}, \boldsymbol{\Delta}_{t|n}, \boldsymbol{\gamma}_{t|n}, \boldsymbol{\Gamma}_{t|n}), \tag{3.13}$$

where $\boldsymbol{\xi}_{t|n} = \boldsymbol{\xi}_{[t]}$, $\boldsymbol{\Omega}_{t|n} = \boldsymbol{\Omega}_{[t,t]}$, $\boldsymbol{\gamma}_{t|n} = \boldsymbol{\gamma}_{1:n|n}$, $\boldsymbol{\Gamma}_{t|n} = \boldsymbol{\Gamma}_{1:n|n}$ *and* $\boldsymbol{\Delta}_{t|n} = \boldsymbol{\Delta}_{1:n|n[t]}$ *denotes the* t*-th block of* p *rows in* $\boldsymbol{\Delta}_{1:n|n}$*. When* $t = n$*, (3.13) gives the filtering distribution at time* n*.*

Another important consequence of Theorem 3.4 is the availability of a closed-form expression for the marginal likelihood $p(\mathbf{y}_{1:n})$, which is provided in Corollary 3.6.

Corollary 3.6. *Under model* (3.1)–(3.2)*, the marginal likelihood has the form* $p(\mathbf{y}_{1:n}) = \Phi_{m \cdot n}\left(\boldsymbol{\gamma}_{1:n|n}; \boldsymbol{\Gamma}_{1:n|n}\right)$*, with* $\boldsymbol{\gamma}_{1:n|n}$ *and* $\boldsymbol{\Gamma}_{1:n|n}$ *as in Theorem 3.4.*

The above result can be useful in several contexts, including empirical Bayes and estimation of unknown system parameters via maximization of the marginal likelihood.

3.4 Inference via Monte Carlo Methods

As discussed in Sections 3.2 and 3.3, inference without sampling from (3.9)–(3.10) or (3.12) is, theoretically, possible. Indeed, since the SUN densities of the filtering, predictive and smoothing distributions are available from Theorems 3.2 and 3.4, the main functionals of interest can be computed either via closed-form expressions (Arellano-Valle and Azzalini, 2006; Azzalini and Bacchieri, 2010; Gupta et al., 2013; Azzalini and Capitanio, 2014) or by relying on numerical integration. However, these strategies require multiple evaluations of multivariate Gaussian cumulative distribution functions. Hence, they tend to be impractical as t increases or when the focus is on complex functionals.

In these situations, Monte Carlo integration provides a tractable solution which allows accurate evaluation of generic functionals $\mathrm{E}[g(\boldsymbol{\beta}_t) \mid \mathbf{y}_{1:t}]$, $\mathrm{E}[g(\boldsymbol{\beta}_{t+1}) \mid \mathbf{y}_{1:t}]$ and $\mathrm{E}[g(\boldsymbol{\beta}_{1:n}) \mid \mathbf{y}_{1:n}]$ for the filtering, predictive and smoothing distribution via

$$\frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\beta}_{t|t}^{(r)}), \quad \frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\beta}_{t+1|t}^{(r)}), \quad \text{and} \quad \frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\beta}_{1:n|n}^{(r)}),$$

where $\boldsymbol{\beta}_{t|t}^{(1)}, \ldots, \boldsymbol{\beta}_{t|t}^{(R)}$, $\boldsymbol{\beta}_{t+1|t}^{(1)}, \ldots, \boldsymbol{\beta}_{t+1|t}^{(R)}$ and $\boldsymbol{\beta}_{1:n|n}^{(1)}, \ldots, \boldsymbol{\beta}_{1:n|n}^{(R)}$ denote random samples from $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\beta}_{t+1} \mid \mathbf{y}_{1:t})$ and $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n})$, respectively. For example, the observations predictive distribution can be computed as $\sum_{r=1}^{R} \Phi_m(\mathbf{B}_{t+1}\mathbf{F}_{t+1}\boldsymbol{\beta}_{t+1|t}^{(r)}; \mathbf{B}_{t+1}\mathbf{V}_{t+1}\mathbf{B}_{t+1})/R$ if the evaluation of (3.11) is computationally demanding.

Clearly, to be implemented, the above approach requires an efficient strategy to sample from (3.9)–(3.10) and (3.12). Exploiting the SUN properties and recent results in Botev (2017), an algorithm to draw independent and identically distributed samples from the exact SUN distributions in (3.9)–(3.10) and (3.12) is developed in Section 3.4.1. As illustrated in Section 3.5, this technique is more accurate than state-of-the-art computational methods and can be efficiently implemented in a variety of small-to-moderate dimensional time series. In Section 3.4.2 we develop, instead, a scalable sequential Monte Carlo scheme for high dimensional settings, which has optimality properties.

3.4.1 Independent and Identically Distributed Sampling

As discussed in Section 3.1, MCMC and sequential Monte Carlo methodologies to sample from $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\beta}_{t+1} \mid \mathbf{y}_{1:t})$ and $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n})$ are available. However, the optimal solution, when possible, is to rely on independent and identically distributed (i.i.d.) samples. Here, we develop a Monte Carlo algorithm to address this goal with a main focus on the smoothing distribution, and discuss immediate modifications to allow sampling also in the filtering and predictive case. Indeed, Monte Carlo inference is particularly suitable in batch settings, although, as discussed later, the proposed routine is useful, in practice, also when the focus is on filtering and predictive distributions, since i.i.d. samples are simulated rapidly, for each t, in small-to-moderate dimensional time series.

Algorithm 4: Independent and identically distributed sampling from $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n})$

1. Sample $\mathbf{U}_{0\ 1:n|n}^{(1)}, \ldots, \mathbf{U}_{0\ 1:n|n}^{(R)}$ independently from a $\mathrm{N}_{p\cdot n}(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{1:n|n} - \boldsymbol{\Delta}_{1:n|n}\boldsymbol{\Gamma}_{1:n|n}^{-1}\boldsymbol{\Delta}_{1:n|n}^{\intercal})$.

2. Sample $\mathbf{U}_{1\ 1:n|n}^{(1)}, \ldots, \mathbf{U}_{1\ 1:n|n}^{(R)}$ independently from a $\mathrm{N}_{m\cdot n}(\mathbf{0}, \boldsymbol{\Gamma}_{1:n|n})$ truncated below $-\boldsymbol{\gamma}_{1:n|n}$.

3. Compute $\boldsymbol{\beta}_{1:n|n}^{(1)}, \ldots, \boldsymbol{\beta}_{1:n|n}^{(R)}$ via $\boldsymbol{\beta}_{1:n|n}^{(r)} = \boldsymbol{\xi}_{1:n|n} + \boldsymbol{\omega}_{1:n|n}(\mathbf{U}_{0\ 1:n|n}^{(r)} + \boldsymbol{\Delta}_{1:n|n}\boldsymbol{\Gamma}_{1:n|n}^{-1}\mathbf{U}_{1\ 1:n|n}^{(r)})$ for each r.

Exploiting the closed-form expression of the smoothing distribution in Theorem 3.4 and the additive representation (3.7) of the SUN, independent realizations $\boldsymbol{\beta}_{1:n|n}^{(1)}, \ldots, \boldsymbol{\beta}_{1:n|n}^{(R)}$ from the smoothing distribution (3.12) can be obtained via a linear combination between independent samples from $(p \cdot n)$-variate Gaussians and $(m \cdot n)$-variate truncated normals. Algorithm 4 provides the pseudo-code for this novel routine, whose outputs are i.i.d. samples from the joint smoothing distribution. Here, the most computationally intensive step is the sampling from the multivariate truncated normal. In fact, although efficient Hamiltonian Monte Carlo solutions are available (Pakman and Paninski, 2014), these strategies do not provide independent samples. More recently, an accept-reject method based on minimax tilting has been proposed by Botev (2017) to improve the acceptance rate of classical rejection sampling, while avoiding convergence and mixing issues of MCMC. Such a routine is available in the `R` library `TruncatedNormal` and allows efficient sampling from multivariate truncated normals having a dimension of few hundreds, thereby providing effective Monte Carlo inference via Algorithm 4 in small-to-moderate dimensional time series.

Clearly, the availability of i.i.d. sampling schemes from the smoothing distribution overcomes the need of MCMC methods and particle smoothers. The first set of strategies face mixing or time-inefficiency issues, especially for imbalanced binary datasets (Johndrow et al., 2019), whereas the second class of routines tend to be computationally intensive and subject to particles degeneracy (Doucet and Johansen, 2009).

When the focus is on Monte Carlo inference for the marginal smoothing distribution $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:n})$ at a specific time t, Algorithm 4 requires minor adaptations relying again on the additive representation of the SUN in equation (3.13), under similar arguments considered for the joint smoothing setting. This latter routine can be also used to sample from the filtering distribution by applying such a scheme with $n = t$ to obtain i.i.d. samples $\boldsymbol{\beta}_{t|t}^{(1)}, \ldots, \boldsymbol{\beta}_{t|t}^{(R)}$ from $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$. Based on these realizations, i.i.d. samples from the predictive distribution can be simply generated via direct application of equation (3.2) to obtain $\boldsymbol{\beta}_{t+1|t}^{(1)} = \mathbf{G}_{t+1}\boldsymbol{\beta}_{t|t}^{(1)} + \boldsymbol{\varepsilon}_{t+1}^{(1)}, \ldots, \boldsymbol{\beta}_{t+1|t}^{(R)} = \mathbf{G}_{t+1}\boldsymbol{\beta}_{t|t}^{(R)} + \boldsymbol{\varepsilon}_{t+1}^{(R)}$, with $\boldsymbol{\varepsilon}_{t+1}^{(1)}, \ldots, \boldsymbol{\varepsilon}_{t+1}^{(R)}$ denoting independent samples from a $\mathrm{N}_p(\mathbf{0}, \mathbf{W}_{t+1})$. Therefore, efficient Monte Carlo inference in small-to-moderate dimensional dynamic probit models is possible also for the filtering and predictive distributions.

3.4.2 Optimal Particle Filtering

When the dimension of the dynamic probit model (3.1)–(3.2) increases, sampling from multivariate truncated Gaussians in Algorithm 4 can face computational bottlenecks (Botev, 2017). This is particularly likely to occur in series monitored on a fine time grid. Indeed, in several applications, the number of time series m is small-to-moderate, whereas the length of the time window can be large. To address this issue and allow scalable online inference for filtering and prediction also in large t settings, we propose a particle filter which exploits the SUN results to obtain optimality properties.

The proposed algorithm is in the class of sequential importance sampling-resampling (SISR) algorithms which provide default strategies in particle filtering (e.g., Doucet et al., 2000, 2001; Durbin and Koopman, 2012). For each time t, these routines sample R trajectories $\boldsymbol{\beta}^{(1)}_{1:t|t}, \ldots, \boldsymbol{\beta}^{(R)}_{1:t|t}$, known as *particles*, conditioned on those produced at $t-1$, by iterating, in time, between the two steps summarized below.

Importance sampling. Let $\boldsymbol{\beta}^{(1)}_{1:t-1|t-1}, \ldots, \boldsymbol{\beta}^{(R)}_{1:t-1|t-1}$ be the particles' trajectories at $t-1$, and denote with $\pi(\boldsymbol{\beta}_{t|t} \mid \boldsymbol{\beta}_{1:t-1}, \mathbf{y}_{1:t})$ the proposal. Then, for each $r = 1, \ldots, R$,

(a) Sample $\bar{\boldsymbol{\beta}}^{(r)}_{t|t}$ from $\pi(\boldsymbol{\beta}_{t|t} \mid \boldsymbol{\beta}^{(r)}_{1:t-1|1:t-1}, \mathbf{y}_{1:t})$ and set $\bar{\boldsymbol{\beta}}^{(r)}_{1:t|t} = (\boldsymbol{\beta}^{(r)\intercal}_{1:t-1|t-1}, \bar{\boldsymbol{\beta}}^{(r)\intercal}_{t|t})^\intercal$.

(b) Compute

$$w^{(r)}_t = w_t(\bar{\boldsymbol{\beta}}^{(r)}_{1:t|t}) \propto p(\mathbf{y}_t \mid \bar{\boldsymbol{\beta}}^{(r)}_{t|t}) \frac{p(\bar{\boldsymbol{\beta}}^{(r)}_{t|t} \mid \boldsymbol{\beta}^{(r)}_{t-1|t-1})}{\pi(\bar{\boldsymbol{\beta}}^{(r)}_{t|t} \mid \boldsymbol{\beta}^{(r)}_{1:t-1|1:t-1}, \mathbf{y}_{1:t})}$$

and normalize these weights to ensure that their sum is 1.

Resampling. For $r = 1, \ldots, R$, sample new particles $\boldsymbol{\beta}^{(1)}_{1:t|t}, \ldots, \boldsymbol{\beta}^{(R)}_{1:t|t}$ from the discrete distribution $\sum_{l=1}^{R} w^{(l)}_t \delta_{\bar{\boldsymbol{\beta}}^{(l)}_{1:t|t}}$.

Based on these particles, functionals of the filtering distribution $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$ can be computed exploiting the terminal values $\boldsymbol{\beta}^{(1)}_{t|t}, \ldots, \boldsymbol{\beta}^{(R)}_{t|t}$ of each trajectory $\boldsymbol{\beta}^{(1)}_{1:t|t}, \ldots, \boldsymbol{\beta}^{(R)}_{1:t|t}$.

As is clear from the above steps, the performance of SISR relies on the proposal $\pi(\boldsymbol{\beta}_{t|t} \mid \boldsymbol{\beta}_{1:t-1}, \mathbf{y}_{1:t})$. This importance function should allow tractable sampling along with efficient evaluation of the importance weights, and should be also carefully specified to propose effective candidate samples. Recalling Doucet et al. (2000), the optimal importance density is $\pi(\boldsymbol{\beta}_{t|t} \mid \boldsymbol{\beta}_{1:t-1}, \mathbf{y}_{1:t}) = p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}, \mathbf{y}_t)$ with weights $w_t(\boldsymbol{\beta}_{1:t}) \propto p(\mathbf{y}_t \mid \boldsymbol{\beta}_{t-1})$. Indeed, this choice minimizes the variance of the importance weights, thereby limiting degeneracy issues and improving mixing. Unfortunately, in several dynamic models, tractable sampling from $p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}, \mathbf{y}_t)$ and direct calculation of $p(\mathbf{y}_t \mid \boldsymbol{\beta}_{t-1})$ is not possible (Doucet et al., 2000). As outlined in Corollary 3.7, this is not the case for multivariate dynamic probit models. In particular, as a direct consequence of Theorem 3.2 and of the closure

Algorithm 5: Optimal particle filter to sample from $p(\boldsymbol{\beta}_t|\mathbf{y}_{1:t})$, for $t = 1, \dots, n$

for t *from* 1 *to* n **do**
  for r *from* 1 *to* R **do**
    1. Propose a value $\bar{\boldsymbol{\beta}}_{t|t}^{(r)}$ by sampling from (3.14) conditioned on $\boldsymbol{\beta}_{t-1} = \boldsymbol{\beta}_{t-1|t-1}^{(r)}$, via
    1.1. Sample $\mathbf{U}_{0\ t|t}^{(r)}$ from a $\mathrm{N}_p(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{t|t,t-1} - \boldsymbol{\Delta}_{t|t,t-1}\boldsymbol{\Gamma}_{t|t,t-1}^{-1}\boldsymbol{\Delta}_{t|t,t-1}^{\intercal})$.
    1.2. Sample $\mathbf{U}_{1\ t|t}^{(r)}$ from a $\mathrm{N}_m(\mathbf{0}, \boldsymbol{\Gamma}_{t|t,t-1})$ truncated below $-\boldsymbol{\gamma}_{t|t,t-1}^{(r)} = -\mathbf{c}_t^{-1}\mathbf{B}_t\mathbf{F}_t\mathbf{G}_t\boldsymbol{\beta}_{t-1|t-1}^{(r)}$.
    1.3. Compute $\bar{\boldsymbol{\beta}}_{t|t}^{(r)} = \mathbf{G}_t\boldsymbol{\beta}_{t-1|t-1}^{(r)} + \boldsymbol{\omega}_{t|t,t-1}(\mathbf{U}_{0\ t|t}^{(r)} + \boldsymbol{\Delta}_{t|t,t-1}\boldsymbol{\Gamma}_{t|t,t-1}^{-1}\mathbf{U}_{1\ t|t}^{(r)})$.
    2. Calculate the associated importance weight $w_t^{(r)}$ via (3.15) and normalize them.
  3. Obtain $\boldsymbol{\beta}_{t|t}^{(1)}, \dots, \boldsymbol{\beta}_{t|t}^{(R)}$ by resampling from $\bar{\boldsymbol{\beta}}_{t|t}^{(1)}, \dots, \bar{\boldsymbol{\beta}}_{t|t}^{(R)}$ with weights $w_t^{(1)}, \dots, w_t^{(R)}$.

properties of the SUN, sampling from $p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}, \mathbf{y}_t)$ is straightforward and $p(\mathbf{y}_t \mid \boldsymbol{\beta}_{t-1})$ has a simple expression.

Corollary 3.7. *Under model* (3.1)–(3.2)*, the following results hold for each* $t = 1, \dots, n$*.*

$$(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}, \mathbf{y}_t) \sim \mathrm{SUN}_{p,m}(\boldsymbol{\xi}_{t|t,t-1}, \boldsymbol{\Omega}_{t|t,t-1}, \boldsymbol{\Delta}_{t|t,t-1}, \boldsymbol{\gamma}_{t|t,t-1}, \boldsymbol{\Gamma}_{t|t,t-1}), \tag{3.14}$$

$$p(\mathbf{y}_t \mid \boldsymbol{\beta}_{t-1}) = \Phi_m(\boldsymbol{\gamma}_{t|t,t-1}; \boldsymbol{\Gamma}_{t|t,t-1}), \tag{3.15}$$

with parameters $\boldsymbol{\xi}_{t|t,t-1} = \mathbf{G}_t\boldsymbol{\beta}_{t-1}$, $\boldsymbol{\Omega}_{t|t,t-1} = \mathbf{W}_t$, $\boldsymbol{\Delta}_{t|t,t-1} = \bar{\boldsymbol{\Omega}}_{t|t,t-1}\boldsymbol{\omega}_{t|t,t-1}\mathbf{F}_t^{\intercal}\mathbf{B}_t\mathbf{c}_t^{-1}$, $\boldsymbol{\gamma}_{t|t,t-1} = \mathbf{c}_t^{-1}\mathbf{B}_t\mathbf{F}_t\boldsymbol{\xi}_{t|t,t-1}$, $\boldsymbol{\Gamma}_{t|t,t-1} = \mathbf{c}_t^{-1}\mathbf{B}_t\left(\mathbf{F}_t\boldsymbol{\Omega}_{t|t,t-1}\mathbf{F}_t^{\intercal} + \mathbf{V}_t\right)\mathbf{B}_t\mathbf{c}_t^{-1}$*, having set* $\mathbf{c}_t = \left[(\mathbf{F}_t\boldsymbol{\Omega}_{t|t,t-1}\mathbf{F}_t^{\intercal} + \mathbf{V}_t) \circ \mathbf{I}_m\right]^{1/2}$.

Algorithm 5 illustrates the pseudo-code of the proposed optimal filter, which exploits the additive representation of the SUN and Corollary 3.7. Comparing Algorithms 4 and 5 it can be noticed that now the computational complexity of the different steps does not depend on t, thus facilitating scalable sequential inference in large t studies. Samples from the predictive distribution can be obtained from those of the filtering as in Section 3.4.1.

3.5 Illustration on Financial Time Series

We study the performance of the methods in Sections 3.3 and 3.4 on a dynamic probit regression for the daily opening directions of the French CAC40 stock market index from January 4th, 2018 to March 29th, 2019. Consistent with this focus, the variable y_t is on a binary scale, with $y_t = 1$ if the opening value of the CAC40 on day t is greater than the corresponding closing value in the previous day, and $y_t = 0$ otherwise. Financial applications of this type have been a source of particular interest in past and recent years (e.g., Kim and Han, 2000; Kara et al., 2011; Atkins et al., 2018), with common approaches combining a wide variety of technical indicators and news information to predict stock

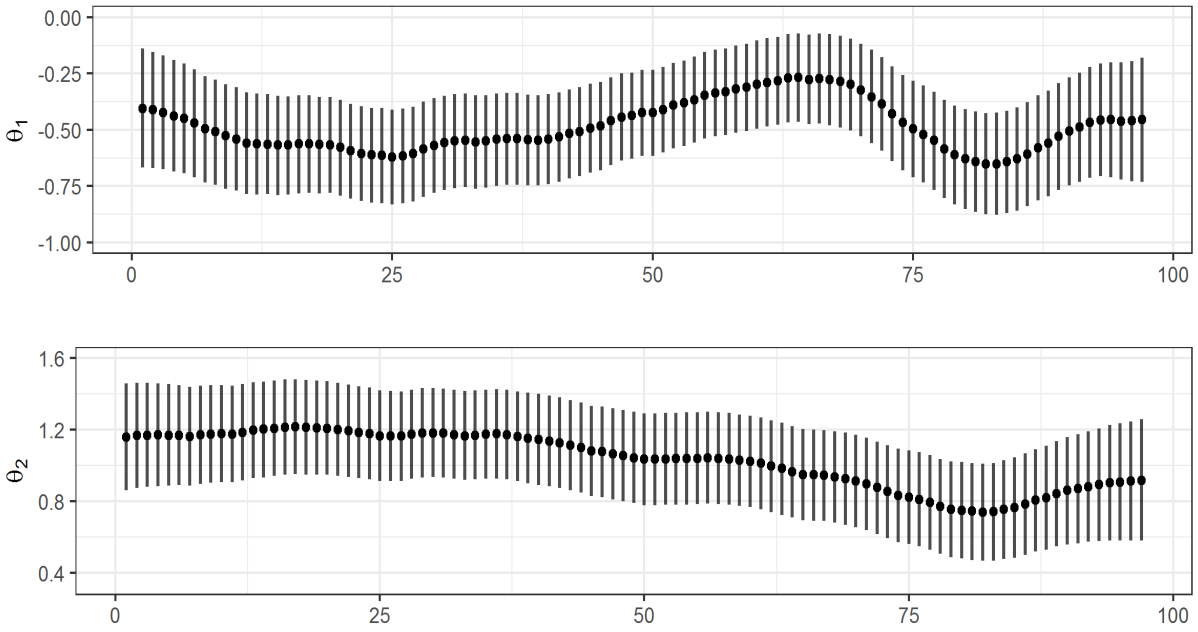

Figure 3.3: Pointwise median and interquartile range for the smoothing distributions of β_{1t} and β_{2t} under the dynamic probit regression in (3.16), for the time window from January 4th, 2018 to May 31st, 2018.

markets directions via complex machine learning methods. Here, we show how a similar predictive performance can be obtained via a simple and interpretable dynamic probit regression for y_t, that combines past information on the opening directions of CAC40 with those of the NIKKEI225, regarded as binary covariates x_t with dynamic coefficients. Since the Japanese market opens before the French one, x_t is available before y_t and, hence, provides a valid predictor for each day t.

Recalling the above discussion and leveraging default specifications in these settings (e.g., Soyer and Sung, 2013), we rely on a dynamic probit regression for y_t with two independent random walk processes for the coefficients $\boldsymbol{\beta}_t = (\beta_{1t}, \beta_{2t})^\intercal$. Letting $\mathbf{F}_t = (1, x_t)$ and $\text{pr}(y_t = 1 \mid \boldsymbol{\beta}_t) = \Phi(\beta_{1t} + \beta_{2t}x_t; 1)$, such a model can be expressed as in equation (3.1) via

$$\begin{aligned} p(y_t \mid \boldsymbol{\beta}_t) &= \Phi[(2y_t - 1)\mathbf{F}_t\boldsymbol{\beta}_t; 1], \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} \text{N}_2(\mathbf{0}, \mathbf{W}), \quad t = 1, \dots n, \quad \boldsymbol{\beta}_0 \sim \text{N}_2(\mathbf{a}_0, \mathbf{P}_0), \end{aligned} \tag{3.16}$$

where $\mathbf{W}$ is a time-invariant diagonal matrix. In (3.16), the element β_{1t} of $\boldsymbol{\beta}_t$ measures the trend in the directions of the CAC40 when the NIKKEI225 has a negative opening on day t, whereas β_{2t} characterizes the shift in such a trend if the opening of the NIKKEI225 index is positive, thereby providing an interpretable probit model for y_t, with dynamic coefficients.

To evaluate performance in smoothing, filtering and prediction, we consider a situation

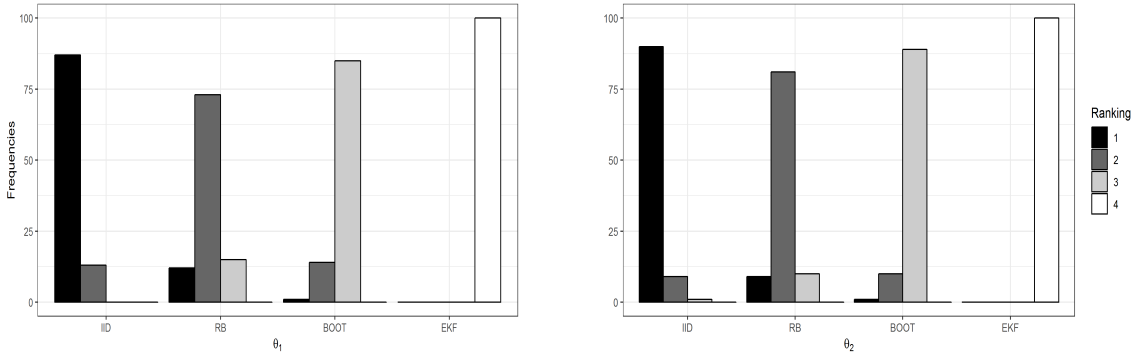

Figure 3.4: Ranking of the four sampling schemes in 100 replicated experiments according to the Wasserstein distance between the empirical smoothing distribution computed, at time $t = 97$, from 10^5 particles and the one obtained by direct evaluation of the exact density (3.10) on two grids of 2000 equally spaced values in $[-2.5, 1.5]$ and $[-1.5, 3]$ for β_{1t} and β_{2t}, respectively, with $t = 97$.

in which the analysis starts on May 31st, 2018, with daily batch data available for the time window from January 4th, 2018 to May 31st, 2018, and online observations streaming in from June 1st, 2018 to March 29th, 2019. This setting motivates smoothing techniques for the first $t = 1, \dots, 97$ times and online filters for the subsequent $t = 98, \dots, 299$ days.

Figure 3.3 shows the pointwise median and interquartile range of the smoothing distribution for β_{1t} and β_{2t}, $t = 1, \dots, 97$, based on 10^5 samples from Algorithm 4. To implement such a routine, we set $\mathbf{a}_0 = (0, 0)^\intercal$ and $\mathbf{P}_0 = \text{diag}(3, 3)$ following the guidelines in Gelman et al. (2008) and Chopin and Ridgway (2017) for probit regression. The states variances in the diagonal matrix $\mathbf{W}$ are instead set equal to 0.01 as suggested by a graphical search of the maximum for the marginal likelihood computed under different combinations of $(\mathbf{W}_{11}, \mathbf{W}_{22})$ via the analytical formula in Corollary 3.6.

As shown in Figure 3.3, the dynamic states β_{1t} and β_{2t} tend to concentrate around negative and positive values, respectively, for the entire smoothing window, thus highlighting a general concordance between CAC40 and NIKKEI225 opening patterns. However, the strength of this association varies in time, supporting our proposed dynamic probit over static specifications. For example, it is possible to observe a decay in β_{1t} and β_{2t} on April–May, 2018 which reduces the association among CAC40 and NIKKEI225, while inducing a general negative trend for the opening directions of the French market. Such a result could be due to the overall instability in the Eurozone on April–May, 2018 caused by the uncertainty after the Italian and British elections during those months.

To clarify the computational improvements provide by the methods in Section 3.4.1, we also compare, in Figure 3.4 and in Table 3.1, their performance against the competing strategies mentioned in Section 3.1. Here, the focus is on the marginal smoothing distribution of β_{1t} and β_{2t} at the last day among those available for batch smoothing. Such a distribution of interest coincides with the filtering at time $t = 97$, thereby allowing the

State	IID	RB	BOOT	EKF
β_{1t} at time $t = 97$	**0.00173**	0.00331	0.00670	0.01845
β_{2t} at time $t = 97$	**0.00221**	0.00428	0.01010	0.06245

Table 3.1: For each sampling scheme, Wasserstein distance, averaged across 100 experiments, between the empirical smoothing distribution computed, at time $t = 97$, from 10^5 particles and the one obtained by direct evaluation of the exact density (3.10) on two grids of 2000 equally spaced values in $[-2.5, 1.5]$ and $[-1.5, 3]$ for β_{1t} and β_{2t}, respectively, with $t = 97$. The lowest distance for each state is bolded.

implementation of the filters discussed in Section 3.1, to evaluate performance both in terms of smoothing and filtering. The competing methods include the extended Kalman filter (Uhlmann, 1992), the bootstrap particle filter (Gordon et al., 1993) and the Rao-Blackwellized sequential Monte Carlo by Andrieu and Doucet (2002) which leverages the hierarchical representation (3.3)–(3.5) of model (3.1)–(3.2). Although being a popular solution in routine implementations, the extended Kalman filter relies on a quadratic approximation of the probit log-likelihood which leads to a Gaussian filtering distribution, thereby affecting the quality of online learning when imbalances in the data induce skewness. The bootstrap particle filter (Gordon et al., 1993) is, instead, motivated by the apparent absence of a tractable optimal proposal distribution $p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}^{(r)}_{t-1|t-1}, \mathbf{y}_t)$ (Doucet et al., 2000) and, therefore, proposes values from $p(\boldsymbol{\beta}_t|\boldsymbol{\beta}^{(r)}_{t-1|t-1})$. Also Rao-Blackwellized sequential Monte Carlo (Andrieu and Doucet, 2002) aims at providing an alternative particle filter, which addresses the apparent unavailability of an analytical expression for the optimal proposal and the corresponding importance weights. The authors overcome this key issue by proposing a sequential Monte Carlo strategy for the Rao-Blackwellized filtering distribution $p(\mathbf{z}_t \mid \mathbf{y}_{1:t})$ of the partially observed Gaussian data $\mathbf{z}_t$ in model (3.3)–(3.5) and compute, for each trajectory $\mathbf{z}^{(r)}_{1:t}$, relevant moments of $p(\boldsymbol{\beta}_t \mid \mathbf{z}^{(r)}_{1:t})$ via classical Kalman filter updates—applied to model (3.4)–(3.5)—which are then averaged across particles to obtain Monte Carlo estimates for the moments of $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$.

Although the above methods provide state-of-the-art solutions, the proposed strategies are motivated by the apparent absence of a closed-form filter for (3.1)–(3.2), which is, in fact, available according to our results in Section 3.3. Figure 3.4 and Table 3.1 highlight to what extent this novel finding improves the existing methods. More specifically, Figure 3.4 compares the rankings of the different sampling schemes, in 100 replicated experiments, according to the Wasserstein distances (e.g., Villani, 2008) between the empirical smoothing distribution induced by the particles generated from each sampling method under analysis and the one obtained by direct evaluation of the exact density (3.10) on an appropriate grid. Table 3.1 shows, instead, these distances averaged across the 100 replicated experiments. For the sake of clarity, with a little abuse of terminology, the term *particle* is used to denote both the samples of the sequential Monte Carlo methods

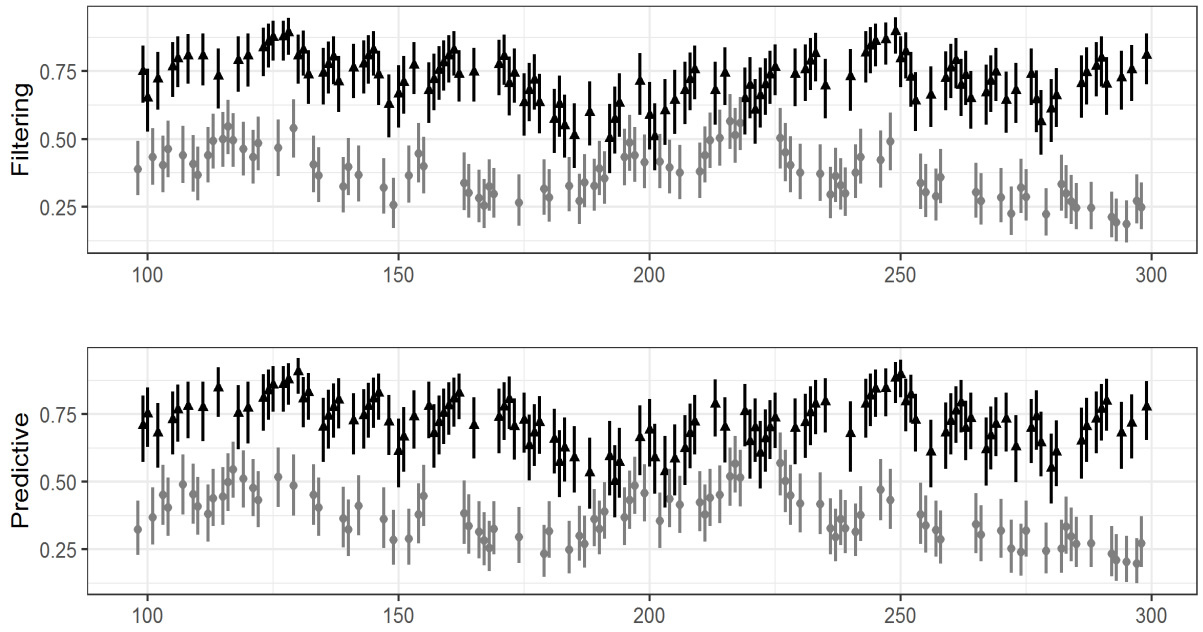

Figure 3.5: Median and interquartile range of the filtering and predictive distributions for $\Phi(\beta_{1t} + x_t\beta_{2t}; 1)$ computed from 10^5 particles produced by the optimal particle filter in Algorithm 5. Black and grey segments denote days in which $x_t = 1$ and $x_t = 0$, respectively.

and those obtained under i.i.d. sampling from the SUN. The Wasserstein distance is computed via the `R` function `wasserstein1d`. Note also that, although the extended Kalman filter and the Rao-Blackwellized sequential Monte Carlo focus, mostly, on the first two central moments of $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t})$, these strategies can be adapted to draw samples from an approximation of the marginal smoothing density.

Figure 3.4 confirms that the sampling scheme in Section 3.4.1 over-performs all the competitors, since its ranking is 1 in most of the 100 experiments. The averaged Wasserstein distances in Table 3.1 yield the same conclusion. Such a result is due to the fact that the extended Kalman filter relies on an approximation of the filtering distribution, whereas, unlike the proposed exact sampler, the bootstrap and the Rao-Blackwellized particle filters consider sub-optimal dependent sampling strategies. Not surprisingly, the Rao-Blackwellized particle filter is the second best choice. Nonetheless, as expected, exact i.i.d. sampling remains the optimal solution and provides a viable strategy in any small-to-moderate study.

Motivated by the accurate performance of the Monte Carlo methods based on SUN results, we also apply the optimal particle filter in Algorithm 5 to provide scalable online filtering and prediction for model (3.16) from June 1st, 2018 to March 29th, 2019. Following the idea of sequential inference, the particles are initialized with the marginal smoothing distribution of May 31, 2018 from the batch analysis. Figure 3.5 outlines median and interquartile range for the filtering and predictive distribution of the probability

that the CAC40 index has a positive opening in each day of the window considered for online inference. These two distributions can be easily obtained by applying the function $\Phi(\beta_{1t} + x_t\beta_{2t}; 1)$ to the particles of the states filtering and predictive distribution. In line with Figure 3.3, a positive opening of the NIKKEI225 provides, in general, an high estimate for the probability that $y_t = 1$, whereas a negative opening tends to favor the event $y_t = 0$. However, the strength of this result evolves over time with some periods showing less evident shifts in the probabilities process when x_t changes from 1 to 0. One-step-ahed prediction, leveraging the samples of the predictive distribution for the probability process, led to a correct classification rate of 66.34% which is comparable to those obtained under more complex procedures combining a wide variety of input information to predict stock markets directions via state-of-the-art machine learning methods (e.g., Kim and Han, 2000; Kara et al., 2011; Atkins et al., 2018).

3.6 Discussion

This chapter shows that the filtering, predictive and smoothing distributions in a dynamic probit model for multivariate binary data have a SUN kernel and the associated parameters can be computed via tractable expressions. As discussed in Sections 3.3–3.5, this result provides advances in online inference for dynamic binary data and facilitates the implementation of tractable methods to draw i.i.d. samples from the exact filtering, predictive and smoothing distributions, thus allowing improved Monte Carlo inference in small-to-moderate time series. High-dimensional filtering can be, instead, implemented via a scalable sequential Monte Carlo which exploits SUN properties to provide a particle filter with optimal proposal.

These results motivate additional future research. For instance, a relevant direction is to generalize the derivations in Section 3.3 to dynamic tobit, binomial and multinomial probit models, for which closed-form filters are unavailable. Joint filtering and prediction of continuous and binary time series is also of interest (Liu et al., 2009). A natural state-space model for these multivariate data can be obtained by generalizing (3.3)–(3.5) to allow only the subset of Gaussian variables associated with the binary data to be partially observed. Also in this case, closed-form filters are not available. By combining our results in Section 3.3 with the classical Kalman filter for Gaussian state-space models, such a gap could be possibly covered. As discussed in Sections 3.1 and 3.3.2, estimation and inference for possible unknown parameters characterizing the state-space model in (3.1)–(3.2) is another interesting problem which can be addressed by maximizing the marginal likelihood derived in Section 3.3.2. Finally, additional quantitative studies beyond those in Section 3.5 can be useful for obtaining a more comprehensive overview on the performance of our proposed computational methods compared to state-of-the-art strategies.

3.A Appendix: Proofs of the main results

3.A.1 Proof of Lemma 3.1

To prove Lemma 3.1, notice that, by applying the Bayes rule, we obtain $p(\boldsymbol{\beta}_1 \mid \mathbf{y}_1) \propto p(\boldsymbol{\beta}_1)p(\mathbf{y}_1 \mid \boldsymbol{\beta}_1)$, where we have $p(\boldsymbol{\beta}_1) = \phi_p(\boldsymbol{\beta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)$ and $p(\mathbf{y}_1 \mid \boldsymbol{\beta}_1) = \Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\beta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1)$. The expression for $p(\boldsymbol{\beta}_1)$ can be easily obtained by noticing that $\boldsymbol{\beta}_1 = \mathbf{G}_1\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_1$ in (3.2), with $\boldsymbol{\beta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0)$ and $\boldsymbol{\varepsilon}_1 \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_1)$. The form for the probability mass function of $(\mathbf{y}_1 \mid \boldsymbol{\beta}_1)$ is instead a direct consequence of equation (3.1). Hence, combining these expressions and recalling (3.6), it is clear that $p(\boldsymbol{\beta}_1 \mid \mathbf{y}_1)$ is proportional to the density of a SUN with suitably–specified parameters, such that the kernel of (3.6) coincides with $\phi_p(\boldsymbol{\beta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\beta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1)$. In particular, letting $\boldsymbol{\xi}_{1|1} = \mathbf{G}_1\mathbf{a}_0$, $\boldsymbol{\Omega}_{1|1} = \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1$, $\boldsymbol{\Delta}_{1|1} = \bar{\boldsymbol{\Omega}}_{1|1}\boldsymbol{\omega}_{1|1}\mathbf{F}_1^\intercal\mathbf{B}_1\mathbf{s}_1^{-1}$, $\boldsymbol{\gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1}$, and $\boldsymbol{\Gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1\mathbf{s}_1^{-1}$, we have that $\boldsymbol{\gamma}_{1|1} + \boldsymbol{\Delta}_{1|1}^\intercal\bar{\boldsymbol{\Omega}}_{1|1}^{-1}\boldsymbol{\omega}_{1|1}^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\xi}_{1|1}) = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1} + \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1(\boldsymbol{\beta}_1 - \boldsymbol{\xi}_{1|1}) = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\beta}_1$, and, in addition, $\boldsymbol{\Gamma}_{1|1} - \boldsymbol{\Delta}_{1|1}^\intercal\bar{\boldsymbol{\Omega}}_{1|1}^{-1}\boldsymbol{\Delta}_{1|1} = \mathbf{s}_1^{-1}[\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1 - \mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal)\mathbf{B}_1]\mathbf{s}_1^{-1} = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1\mathbf{s}_1^{-1}$, with $\mathbf{s}_1^{-1}$ as in Lemma 3.1. Now, substituting these quantities in the kernel of the SUN density (3.6), we have

$$\begin{aligned}
&\phi_p(\boldsymbol{\beta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi_m(\mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\beta}_1; \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1\mathbf{s}_1^{-1}) \\
&\qquad = \phi_p(\boldsymbol{\beta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\beta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1) \\
&\qquad = p(\boldsymbol{\beta}_1)p(\mathbf{y}_1 \mid \boldsymbol{\beta}_1) \propto p(\boldsymbol{\beta}_1 \mid \mathbf{y}_1),
\end{aligned}$$

thus proving Lemma 3.1. To prove that $\boldsymbol{\Omega}_{1|1}^*$ is a correlation matrix, replace the identity $\mathbf{I}_m$ with $\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1$ in the proof of Theorem 1 by Durante (2019). □

3.A.2 Proof of Theorem 3.2

Recalling (3.2), the proof for $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})$ in (3.9) requires studying the variable $\mathbf{G}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\varepsilon}_t$, given $\mathbf{y}_{1:t-1}$, where

$$(\boldsymbol{\beta}_{t-1} \mid \mathbf{y}_{1:t-1}) \sim \mathrm{SUN}_{p,m\cdot(t-1)}(\boldsymbol{\xi}_{t-1|t-1}, \boldsymbol{\Omega}_{t-1|t-1}, \boldsymbol{\Delta}_{t-1|t-1}, \boldsymbol{\gamma}_{t-1|t-1}, \boldsymbol{\Gamma}_{t-1|t-1})$$

and $\boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t)$, with $\boldsymbol{\varepsilon}_t \perp \mathbf{y}_{1:t-1}$. To address this goal, first note that, by the closure properties of the unified skew-normal under linear transformations (Azzalini and Capitanio, 2014, Chapter 7.1.2), the variable $(\mathbf{G}_t\boldsymbol{\beta}_{t-1} \mid \mathbf{y}_{1:t-1})$ is still a unified skew-normal and has parameters $\mathbf{G}_t\boldsymbol{\xi}_{t-1|t-1}$, $\mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal$, $[(\mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal) \circ \mathbf{I}_p]^{-1/2}\mathbf{G}_t\boldsymbol{\omega}_{t-1|t-1}\boldsymbol{\Delta}_{t-1|t-1}$, $\boldsymbol{\gamma}_{t-1|t-1}$ and $\boldsymbol{\Gamma}_{t-1|t-1}$. Hence, to conclude the proof of equation (3.9), we only need to obtain the distribution of the sum among this variable and the noise $\boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t)$. This can be accomplished by considering the moment generating function of such a sum—as done by Azzalini and Capitanio (2014, Chapter 7.1.2) to prove closure under convolution. Indeed, it is straightforward to notice that the product of the moment generating functions

for $\boldsymbol{\varepsilon}_t$ and $(\mathbf{G}_t\boldsymbol{\beta}_{t-1} \mid \mathbf{y}_{1:t-1})$ leads to the moment generating function of the unified skew-normal variable having parameters $\boldsymbol{\xi}_{t|t-1} = \mathbf{G}_t\boldsymbol{\xi}_{t-1|t-1}$, $\boldsymbol{\Omega}_{t|t-1} = \mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal + \mathbf{W}_t$, $\boldsymbol{\Delta}_{t|t-1} = \boldsymbol{\omega}_{t|t-1}^{-1}\mathbf{G}_t\boldsymbol{\omega}_{t-1|t-1}\boldsymbol{\Delta}_{t-1|t-1}$, $\boldsymbol{\gamma}_{t|t-1} = \boldsymbol{\gamma}_{t-1|t-1}$ and $\boldsymbol{\Gamma}_{t|t-1} = \boldsymbol{\Gamma}_{t-1|t-1}$.

To prove (3.10) note that $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t}) \propto \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})$ coincides with the posterior distribution in a probit model with likelihood $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ and SUN prior $p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})$ from (3.9). Hence, expression (3.10) can be derived from Corollary 4 in Durante (2019), replacing the matrix $\mathbf{I}_m$ in the classical probit likelihood with $\mathbf{B}_t\mathbf{V}_t\mathbf{B}_t$. □

3.A.3 Proof of Corollary 3.3

To prove Corollary 3.3, first notice that $\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})\mathrm{d}\boldsymbol{\beta}_t$ can be re-written as $\Phi_{m\cdot(t-1)}(\boldsymbol{\gamma}_{t|t-1}; \boldsymbol{\Gamma}_{t|t-1})^{-1} \int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})\mathrm{d}\boldsymbol{\beta}_t$ where $K(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1}) = p(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})\Phi_{m\cdot(t-1)}(\boldsymbol{\gamma}_{t|t-1}; \boldsymbol{\Gamma}_{t|t-1})$ is the kernel of the predictive distribution in equation (3.9). Consistent with this expression, Corollary 3.3 follows after noticing that $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})$ is the kernel of the filtering distribution in (3.10), whose normalizing constant $\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\beta}_t \mid \mathbf{y}_{1:t-1})\mathrm{d}\boldsymbol{\beta}_t$ is equal to $\Phi_{m\cdot t}(\boldsymbol{\gamma}_{t|t}; \boldsymbol{\Gamma}_{t|t})$. □

3.A.4 Proof of Theorem 3.4

First notice that $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n}) \propto p(\boldsymbol{\beta}_{1:n})p(\mathbf{y}_{1:n} \mid \boldsymbol{\beta}_{1:n})$. Hence, $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n})$ can be interpreted as the posterior distribution in the Bayesian model having likelihood $p(\mathbf{y}_{1:n} \mid \boldsymbol{\beta}_{1:n})$ and prior $p(\boldsymbol{\beta}_{1:n})$ for the $(p \cdot n) \times 1$ vector $\boldsymbol{\beta}_{1:n} = (\boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_n^\intercal)^\intercal$. As already pointed out in Section 3.3.2, it immediately follows from the model specification (3.2) that $\boldsymbol{\beta}_{1:n} \sim \mathrm{N}_{p\cdot n}(\boldsymbol{\xi}, \boldsymbol{\Omega})$, with $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$ as in Section 3.3.2. The form of $p(\mathbf{y}_{1:n} \mid \boldsymbol{\beta}_{1:n})$ can be instead obtained from (3.1), by noticing that given $\boldsymbol{\beta}_{1:n}$ the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are conditionally independent, thus providing the joint likelihood $p(\mathbf{y}_{1:n} \mid \boldsymbol{\beta}_{1:n}) = \prod_{t=1}^n \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$. Such a quantity can be also expressed as $\Phi_{m\cdot n}(\mathbf{D}\boldsymbol{\beta}_{1:n}; \mathbf{V})$ with $\mathbf{D}$ and $\mathbf{V}$ as in Section 3.3.2. Combining these results, the joint smoothing distribution $p(\boldsymbol{\beta}_{1:n} \mid \mathbf{y}_{1:n})$ is proportional to $\phi_{p\cdot n}(\boldsymbol{\beta}_{1:n} - \boldsymbol{\xi}; \boldsymbol{\Omega})\Phi_{m\cdot n}(\mathbf{D}\boldsymbol{\beta}_{1:n}; \mathbf{V})$, which is the kernel of a unified skew-normal random variable with parameters as in (3.12). □

3.A.5 A.5. Proof of Corollary 3.6

The formula for the marginal likelihood follows easily after noticing that $p(\mathbf{y}_{1:n})$ coincides with the normalizing constant of the joint smoothing distribution. Indeed, $p(\mathbf{y}_{1:n})$ is formally defined as $\int p(\mathbf{y}_{1:n} \mid \boldsymbol{\beta}_{1:n})p(\boldsymbol{\beta}_{1:n})d\boldsymbol{\beta}_{1:n}$. Hence, the integrand function coincides with the kernel of the smoothing density, so that the whole integral is equal to $\Phi_{m\cdot n}(\boldsymbol{\gamma}_{1:n|n}; \boldsymbol{\Gamma}_{1:n|n})$. □

3.A.6 Proof of Corollary 3.7

The proof of Corollary 3.7 is similar to the one of Lemma 3.1. Indeed, the proposal $p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}, \mathbf{y}_t)$ is the posterior distribution in a Bayesian probit regression with likelihood $p(\mathbf{y}_t \mid \boldsymbol{\beta}_t) = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\beta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ and prior $p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}) = \phi_p(\boldsymbol{\beta}_t - \mathbf{G}_t\boldsymbol{\beta}_{t-1}; \mathbf{W}_t)$. To derive the expression of the importance weights in equation (3.15), it suffices to notice that the marginal likelihood $p(\mathbf{y}_t \mid \boldsymbol{\beta}_{t-1})$ coincides with the normalizing constant of the SUN in (3.14). □

Chapter 4

Conjugate Bayes for Multinomial Probit Models

4.1 Introduction

Regression models for categorical data are ubiquitous in various fields of application and play a fundamental role in classification (e.g., Agresti, 2013). Within this framework, the overarching goal is to learn how a vector of class probabilities, be it $\boldsymbol{\pi}(\mathbf{x}_i) = [\pi_1(\mathbf{x}_i), \ldots, \pi_L(\mathbf{x}_i)]^\intercal = [p(y_i = 1 \mid \boldsymbol{\beta}, \mathbf{x}_i), \ldots, p(y_i = L \mid \boldsymbol{\beta}, \mathbf{x}_i)]^\intercal$, changes with a set of p attributes $\mathbf{x}_i$ observed for each statistical unit i, where $\boldsymbol{\beta}$ denotes a vector of parameters controlling the attribute effects. We refer to Maddala (1986); Greene (2003) and Agresti (2013) for an overview of popular formulations to address such a goal, and focus in this chapter on the class of multinomial probit models. Indeed, such a broad set of formulations has gained vast popularity in social science, economics and machine learning applications, among others, due to their natural connection with Gaussian regression models that act as latent predictor–dependent random utilities in a discrete choice setting, and also allow improved interpretability (Hausman and Wise, 1978; Daganzo, 2014). Moreover, expressing predictor–dependent class probabilities via correlated Gaussian latent utilities facilitates improved flexibility, thus avoiding restrictive assumptions, such as the *independence of irrelevant alternatives* (Hausman and Wise, 1978). Such desirable properties have motivated various generalizations of the original formulation proposed by Hausman and Wise (1978), to incorporate class–specific predictor effects (Stern, 1992) and sequential discrete choices (Tutz, 1991), that have also featured successful implementations and extensions in machine learning (Girolami and Rogers, 2006; Rogers and Girolami, 2007; Riihimäki et al., 2013; Johndrow et al., 2013; Agarwal et al., 2014; Kindo et al., 2016).

The above benefits come, however, with computational difficulties in dealing with integrals of multivariate Gaussian densities (Genz, 1992; Horrace, 2005; Chopin, 2011;

Botev, 2017; Genton et al., 2018; Cao et al., 2019). These key challenges have stimulated intensive research both in frequentist and Bayesian settings. In this chapter, we aim to provide theoretical, methodological and computational advances for the second class of approaches to inference. Indeed, while the frequentist methods for estimation, inference and classification in multinomial probit models are relatively well–established (McFadden, 1989; Stern, 1992; Börsch-Supan and Hajivassiliou, 1993; Geweke et al., 1994; Natarajan et al., 2000), state–of–the–art Bayesian solutions rely either on Markov chain Monte Carlo (MCMC) strategies (Albert and Chib, 1993; McCulloch and Rossi, 1994; Nobile, 1998; McCulloch et al., 2000; Albert and Chib, 2001; Chen and Kuo, 2002; Imai and Van Dyk, 2005; Zhang et al., 2006; Burgette and Nordheim, 2012; Johndrow et al., 2013) or on approximations of the posterior (Girolami and Rogers, 2006; Girolami and Zhong, 2007; Riihimäki et al., 2013; Knowles and Minka, 2011). Despite being widely implemented, both solutions still raise open questions in terms of accuracy and computational tractability, especially in large p settings and in imbalanced situations where some classes are relatively less frequent than others. Indeed, as discussed by Chopin and Ridgway (2017); Johndrow et al. (2019); Durante (2019) and in Chapter 2 of the present thesis, such issues arise even in simple univariate probit models. Moreover, MCMC and approximate methods are still sub–optimal relative to situations in which the posterior is analytically available from a tractable class of distributions.

In Sections 4.2–4.3, we prove that the entire class of unified skew–normal (SUN) distributions (Arellano-Valle and Azzalini, 2006)—which encompasses a broad variety of random variables, including classical Gaussian ones, as already pointed out in previous chapters—is conjugate to various multinomial probit models (Hausman and Wise, 1978; Stern, 1992; Tutz, 1991). Such a broad class of prior distributions has been originally developed in seemingly unrelated contexts to introduce skewness in a multivariate Gaussian density via the cumulative distribution function of another Gaussian vector, thus retaining several probabilistic properties of multivariate Gaussian variables (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014). Leveraging such properties, we derive in Section 4.3 closed–form expressions for posterior predictive distributions and marginal likelihoods which can be used for classification, variable selections and estimation of fixed parameters. Evaluation of more complex functionals proceeds instead via improved Monte Carlo methods which, unlike for state–of–the–art MCMC routines, rely on independent and identically distributed samples from the exact posterior, thus avoiding mixing issues and convergence diagnostics. As discussed in Section 4.3.2.1, such an improved strategy deals with multivariate truncated normals and cumulative distribution functions of multivariate Gaussians whose dimension grows with the sample size n. Therefore, the proposed strategy is particularly useful, in practice, in small–to–moderate n situations, and massively improves state–of–the–art solutions in large p studies, a setting which occurs in various applications but is computationally impractical under the

available implementations (Chopin and Ridgway, 2017). To address the scalability issues of the methods proposed in Section 4.3.2.1, we further improve and extend in Section 4.3.2.2 partially–factorized variational methods for univariate probit models (see Chapter 2) to devise novel blocked partially–factorized approximations of the posterior in multinomial probit regression that easily scale to large p and n datasets, and almost perfectly matches the exact posterior, especially when $p > n$. These findings are illustrated in a gastrointestinal lesions application (Mesejo et al., 2016) in Section 4.4. Finally, Section 4.5 presents future directions of research which highlight how these novel results can motivate applied, methodological and computational advances in multinomial probit models. All proofs can be found in Appendix 4.A, and combine properties of multivariate Gaussians and SUN random variables. Early findings on this connection are presented in Durante (2019) with a focus on Bayesian univariate binary probit models. Such results are special cases of our broader derivations which require non–straightforward and novel extensions to incorporate classical multinomial probit models (Hausman and Wise, 1978) and their generalizations (Stern, 1992; Tutz, 1991). Relative to the univariate case, such models rely on more complex latent variable representations, typically based on the maximum of a multivariate vector of latent utilities that usually require a separate treatment relative to the univariate case, as clarified in Section 4.2.

4.2 Multinomial Probit Models

In this section we review three widely–implemented multinomial probit models that cover a large range of applications. These include the original formulation proposed by Hausman and Wise (1978), and two subsequent generalizations which account for class–specific predictor effects (Stern, 1992) and sequential discrete choices (Tutz, 1991). Despite providing different generative mechanisms for the class probability vector $\boldsymbol{\pi}(\mathbf{x}_i) = [\pi_1(\mathbf{x}_i), \ldots, \pi_L(\mathbf{x}_i)]^\intercal$, as discussed in Sections 4.2.1–4.2.3, all these representations rely on latent Gaussian random utilities and their likelihood can be expressed via the cumulative distribution function of a multivariate Gaussian. This facilitates the derivation of the conjugacy results in Section 4.3.

4.2.1 Classical Discrete Choice Multinomial Probit Models

Let us first focus on the classical discrete choice model as originally formulated by Hausman and Wise (1978). Recalling Greene (2003, sec.4 18.2.6), such a representation expresses each class probability $\pi_l(\mathbf{x}_i)$ via a random utility model in which every unit i chooses among L possible alternatives by maximizing a set of latent Gaussian utilities $z_{i1}, \ldots, z_{iL}$ which depend on p–dimensional vectors $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iL}$ of class–specific attributes as perceived by unit i. More specifically, each $\pi_l(\mathbf{x}_i)$ in $\boldsymbol{\pi}(\mathbf{x}_i) = [\pi_1(\mathbf{x}_i), \ldots, \pi_L(\mathbf{x}_i)]^\intercal$ is

expressed as

$$p(y_i = l \mid \boldsymbol{\beta}, \mathbf{x}_i) = p(z_{il} > z_{ik}, \forall k \neq l) = p(\mathbf{x}_{il}^\intercal \boldsymbol{\beta} + \varepsilon_{il} > \mathbf{x}_{ik}^\intercal \boldsymbol{\beta} + \varepsilon_{ik}, \forall k \neq l), \tag{4.1}$$

for each $l = 1, \ldots, L$, where $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iL})^\intercal \sim \mathrm{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$, independently for $i = 1, \ldots, n$; see Greene (2003) for indentifiability restrictions on $\boldsymbol{\Sigma}$. In (4.1), the generic vector $\mathbf{x}_{il} = (x_{il1}, \ldots, x_{ilp})^\intercal$ of predictors has elements x_{ilj} measuring how the jth attribute of the alternative l is perceived by unit i. For instance, in political studies (e.g. Dow and Endersby, 2004), each $\mathbf{x}_{il}$ can include both information on voter i and attributes of candidate l as perceived by voter i. Hence, this specification assumes that to each individual i are associated L vectors of p observed predictors whose linear combinations $\mathbf{x}_{i1}^\intercal \boldsymbol{\beta}, \ldots, \mathbf{x}_{iL}^\intercal \boldsymbol{\beta}$ contribute to defining the class–specific latent utilities $z_{i1}, \ldots, z_{iL}$. Each individual i will then choose the alternative with the highest random utility $z_{il} = \mathbf{x}_{il}^\intercal \boldsymbol{\beta} + \varepsilon_{il}$ which is defined by a deterministic component $\mathbf{x}_{il}^\intercal \boldsymbol{\beta}$ with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$, plus a Gaussian noise ε_{il}. This term accounts for deviations from the deterministic part due to potential unobserved attributes and, as stated in Proposition 4.1, it induces a joint likelihood for the observed responses $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ that coincides with the cumulative distribution function of an $n \cdot (L-1)$–variate Gaussian.

Proposition 4.1. *Let $\mathbf{v}_l$ denote an $L \times 1$ vector having value 1 in position l and 0 elsewhere, for every $l = 1, \ldots, L$. Moreover, for every $l = 1, \ldots, L$, denote with $\mathbf{V}_{[-l]}$ and $\mathbf{X}_{i[-l]}$ the $(L-1) \times L$ and $(L-1) \times p$ matrices whose rows are obtained by stacking vectors $(\mathbf{v}_k - \mathbf{v}_l)^\intercal$ and $(\mathbf{x}_{il} - \mathbf{x}_{ik})^\intercal$, respectively, for $k = 1, \ldots, l-1, l+1, \ldots L$. Then, under model (4.1) with $\boldsymbol{\varepsilon}_i \sim \mathrm{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$, independently for every unit $i = 1, \ldots, n$, we have*

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\beta}, \mathbf{x}_i) = \prod_{i=1}^{n} \Phi_{L-1}(\mathbf{X}_{i[-y_i]}\boldsymbol{\beta}; \mathbf{V}_{[-y_i]}\boldsymbol{\Sigma}\mathbf{V}_{[-y_i]}^\intercal) = \Phi_{n\cdot(L-1)}(\bar{\mathbf{X}}\boldsymbol{\beta}; \boldsymbol{\Lambda}), \tag{4.2}$$

where $\bar{\mathbf{X}}$ is an $n \cdot (L-1) \times p$ block matrix with $(L-1) \times p$ blocks $\bar{\mathbf{X}}_{[i1]} = \mathbf{X}_{i[-y_i]}$, for each $i = 1, \ldots, n$, whereas $\boldsymbol{\Lambda}$ denotes an $n \cdot (L-1) \times n \cdot (L-1)$ block diagonal covariance matrix with $(L-1) \times (L-1)$ diagonal blocks $\boldsymbol{\Lambda}_{[ii]} = \mathbf{V}_{[-y_i]}\boldsymbol{\Sigma}\mathbf{V}_{[-y_i]}^\intercal$, for $i = 1, \ldots, n$. In (4.2), the generic function $\Phi_c(\mathbf{w}; \mathbf{S})$ denotes the cumulative distribution function, evaluated at $\mathbf{w}$, of a c–variate Gaussian with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{S}$.

The above results follow from (4.1) after noticing that $p(y_i = l \mid \boldsymbol{\beta}, \mathbf{x}_i)$ can be written as $p[\varepsilon_{ik} - \varepsilon_{il} < (\mathbf{x}_{il} - \mathbf{x}_{ik})^\intercal \boldsymbol{\beta}, \forall k \neq l] = p(\mathbf{V}_{[-l]}\boldsymbol{\varepsilon}_i < \mathbf{X}_{i[-l]}\boldsymbol{\beta}) = \Phi_{L-1}(\mathbf{X}_{i[-l]}\boldsymbol{\beta}; \mathbf{V}_{[-l]}\boldsymbol{\Sigma}\mathbf{V}_{[-l]}^\intercal)$, where $\boldsymbol{\varepsilon}_i \sim \mathrm{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$ and, hence, $\mathbf{V}_{[-l]}\boldsymbol{\varepsilon}_i \sim \mathrm{N}_{L-1}(\mathbf{0}, \mathbf{V}_{[-l]}\boldsymbol{\Sigma}\mathbf{V}_{[-l]}^\intercal)$. The final equality in (4.2) is instead a direct consequence of the properties of multivariate Gaussian random variables. Indeed, since $\boldsymbol{\Lambda}$ is a block diagonal covariance matrix and $\bar{\mathbf{X}}\boldsymbol{\beta}$ is obtained by stacking sub–vectors $\mathbf{X}_{i[-y_i]}\boldsymbol{\beta}$ for $i = 1, \ldots, n$, it follows that $\Phi_{n\cdot(L-1)}(\bar{\mathbf{X}}\boldsymbol{\beta}; \boldsymbol{\Lambda})$ factorizes as the product of n cumulative distribution functions of $(L-1)$–variate Gaussians.

As previously mentioned, this formulation has been originally developed in social science and economic studies where there is a vector of predictors $\mathbf{x}_{il}$ for each combination of unit i and class l (Hausman and Wise, 1978). This is, however, not always the case in general classification settings. Indeed, in such situations it is more common to observe a single vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$ of p predictors for each statistical unit $i = 1, \ldots, n$ and the focus is on modeling the vector $\boldsymbol{\pi}(\mathbf{x}_i) = [\pi_1(\mathbf{x}_i), \ldots, \pi_L(\mathbf{x}_i)]^\intercal$, to ultimately predict the class y_i of unit i. In Sections 4.2.2 and 4.2.3 we focus on two widely–implemented representations (Stern, 1992; Tutz, 1991), which address this goal, while still relying on Gaussian latent utilities.

4.2.2 Discrete Choice Multinomial Probit Models with Class–Specific Effects

When a single vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$ of p covariates is observed for each unit $i = 1, \ldots, n$, an interpretable and common solution to model differences in the class probabilities within $\boldsymbol{\pi}(\mathbf{x}_i) = [\pi_1(\mathbf{x}_i), \ldots, \pi_L(\mathbf{x}_i)]^\intercal$ is to introduce class–specific predictors effects $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L$ as in Stern (1992), and define again $\pi_l(\mathbf{x}_i)$ as a function of Gaussian utilities $z_{i1}, \ldots, z_{iL}$ via

$$p(y_i = l \mid \boldsymbol{\beta}, \mathbf{x}_i) = p(z_{il} > z_{ik}, \forall k \neq l) = p(\mathbf{x}_i^\intercal \boldsymbol{\beta}_l + \varepsilon_{il} > \mathbf{x}_i^\intercal \boldsymbol{\beta}_k + \varepsilon_{ik}, \forall k \neq l), \tag{4.3}$$

for each $l = 1, \ldots, L$, where $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iL})^\intercal \sim \mathrm{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$, independently for every unit $i = 1, \ldots, n$, and $\boldsymbol{\beta}_L = \mathbf{0}$ for identifiability purposes (Johndrow et al., 2013). Representation (4.3) and its interpretation are closely related to the classical discrete choice multinomial probit model in Section 4.2.1, with the only key exception that the differences in the class–specific latent utilities $z_{i1}, \ldots, z_{iL}$, $i = 1, \ldots, n$, are now driven by changes in the coefficients vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L$, rather than in the predictors' vectors as in (4.1). For instance, recalling the political example discussed in Section 4.2.1, although the age is an attribute specific to voter i, it is reasonable to expect that such a covariate has a different effect in producing the utilities $z_{i1} = \mathbf{x}_i^\intercal \boldsymbol{\beta}_1 + \varepsilon_{i1}, \ldots, z_{iL} = \mathbf{x}_i^\intercal \boldsymbol{\beta}_L + \varepsilon_{iL}$ that voter i assigns to the different candidates $l = 1, \ldots, L$. This property can be included by allowing the coefficient associated with the age attribute to change across classes, thus providing a formulation more similar to classical multinomial logit models (e.g. Greene, 2003), relative to (4.1). As stated in Proposition 4.2, also under this representation the likelihood for the observed responses $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ coincides with the cumulative distribution function of an $n \cdot (L-1)$–variate Gaussian.

Proposition 4.2. *Denote with $\mathbf{v}_l$ the $L \times 1$ vector with value 1 in position l and 0 elsewhere, for each $l = 1, \ldots, L$. Moreover, let $\mathbf{x}_{il} = \bar{\mathbf{v}}_l \otimes \mathbf{x}_i$, where $\bar{\mathbf{v}}_l$ is the $(L-1) \times 1$ vector obtained by the removing the L-th element from $\mathbf{v}_l$, and $\otimes$ denotes the Kronecker*

product. Then, under model (4.3) *with* $\boldsymbol{\varepsilon}_i \sim \mathrm{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$*, independently for each unit* $i = 1, \ldots, n$*, we have*

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\beta}, \mathbf{x}_i) = \prod_{i=1}^{n} \Phi_{L-1}(\mathbf{X}_{i[-y_i]}\boldsymbol{\beta}; \mathbf{V}_{[-y_i]}\boldsymbol{\Sigma}\mathbf{V}_{[-y_i]}^{\intercal}) = \Phi_{n\cdot(L-1)}(\bar{\mathbf{X}}\boldsymbol{\beta}; \boldsymbol{\Lambda}), \tag{4.4}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\intercal}, \ldots, \boldsymbol{\beta}_{L-1}^{\intercal})^{\intercal}$*, whereas* $\mathbf{X}_{i[-y_i]}$*,* $\mathbf{V}_{[-y_i]}$*,* $\bar{\mathbf{X}}$ *and* $\boldsymbol{\Lambda}$ *are defined as in Proposition 4.1, setting* $\mathbf{x}_{il} = \bar{\mathbf{v}}_l \otimes \mathbf{x}_i$ *for each* $i = 1, \ldots, n$ *and* $l = 1, \ldots, L$*.*

Proposition 4.2 follows as a directed consequence of Proposition 4.1, upon noticing that model (4.3) can be re–written as a particular case of model (4.1) with working covariates $\mathbf{x}_{il}$ as defined in Proposition 4.2. Indeed, note that by setting $\mathbf{x}_{il} = \bar{\mathbf{v}}_l \otimes \mathbf{x}_i$, $i = 1, \ldots, n$, $l = 1, \ldots, L$, the class probabilities in (4.3) can be expressed as $p(y_i = l \mid \boldsymbol{\beta}, \mathbf{x}_i) = p(z_{il} > z_{ik}, \forall k \neq l) = p(\mathbf{x}_i^{\intercal}\boldsymbol{\beta}_l + \varepsilon_{il} > \mathbf{x}_i^{\intercal}\boldsymbol{\beta}_k + \varepsilon_{ik}, \forall k \neq l) = p(\mathbf{x}_{il}^{\intercal}\boldsymbol{\beta} + \varepsilon_{il} > \mathbf{x}_{ik}^{\intercal}\boldsymbol{\beta} + \varepsilon_{ik}, \forall k \neq l)$, for $l = 1, \ldots, L$, with $\boldsymbol{\beta}_L = \mathbf{0}$, where the last quantity is the expression for the class probabilities in (4.1).

4.2.3 Sequential Discrete Choice Multinomial Probit Models

Before focusing on prior specification and posterior derivations, we consider also an extension of the sequential discrete choice multinomial probit model studied in Albert and Chib (2001) and originally proposed by Tutz (1991). This model still relies on a set of class–specific latent utilities but is conceptually different from those presented in Section 4.2.1 and 4.2.2, since the choice among the L classes is modeled via a sequence of binary decisions where the generic step l of this sequential decision process is reached if individual i has not chosen classes $1, \ldots, l-1$, and the binary decision at this step will be to either pick class l with probability $p(y_i = l \mid y_i > l-1, \boldsymbol{\beta}, \mathbf{x}_i) = \Phi(\mathbf{x}_i^{\intercal}\boldsymbol{\beta}_l)$ or to consider one of the subsequent alternatives $l+1, \ldots, L$ with complement probability $p(y_i > l \mid y_i > l-1, \boldsymbol{\beta}, \mathbf{x}_i) = 1 - \Phi(\mathbf{x}_i^{\intercal}\boldsymbol{\beta}_l)$. Note that relative to the original formulations in Albert and Chib (2001) and Tutz (1991), here we consider a slightly different reparameterization and also allow the entire vector of coefficients, and not just the intercept, to change with the different labels, thus providing a more general representation. As discussed by Albert and Chib (2001) also this model has a latent utility representation which expresses each $\pi_l(\mathbf{x}_i)$ in $\boldsymbol{\pi}(\mathbf{x}_i) = [\pi_1(\mathbf{x}_i), \ldots, \pi_L(\mathbf{x}_i)]^{\intercal}$ as

$$p(y_i = l \mid \boldsymbol{\beta}, \mathbf{x}_i) = p(z_{il} > 0) \prod_{k=1}^{l-1} p(z_{ik} < 0) = p(\mathbf{x}_i^{\intercal}\boldsymbol{\beta}_l + \varepsilon_{il} > 0) \prod_{k=1}^{l-1} p(\mathbf{x}_i^{\intercal}\boldsymbol{\beta}_k + \varepsilon_{ik} < 0), \tag{4.5}$$

for $l = 1, \ldots, L-1$, and $p(y_i = L \mid \boldsymbol{\beta}, \mathbf{x}_i) = \prod_{k=1}^{L-1} p(\mathbf{x}_i^{\intercal}\boldsymbol{\beta}_k + \varepsilon_{ik} < 0)$, where $\varepsilon_{il} \sim \mathrm{N}(0, 1)$ independently for unit $i = 1, \ldots, n$ and class $l = 1, \ldots, L-1$. Equation (4.5) provides a general representation in which each $z_{il} = \mathbf{x}_i^{\intercal}\boldsymbol{\beta}_l + \varepsilon_{il}$ denotes the utility of choosing

alternative l against the subsequent ones $l+1,\ldots,L$, given that the classes $1,\ldots,l-1$ have been discarded in the previous steps of the sequential decision process. Proposition 4.3 shows that, although conceptually different from the models in Sections 4.2.1 and 4.2.2, also such a formulation admits a very similar expression for the joint likelihood of data $\mathbf{y}=(y_1,\ldots,y_n)^\intercal$.

Proposition 4.3. *Define* $\bar{\mathbf{y}}_i=(\mathbf{0}_{y_i-1}^\intercal,1)^\intercal$ *if* $y_i\leq L-1$, *and* $\bar{\mathbf{y}}_i=\mathbf{0}_{L-1}$ *if* $y_i=L$, *where the generic* $\mathbf{0}_c$ *is a* $c\times 1$ *vector of zeroes. Moreover, let* $\bar{n}=n_1+\cdots+n_n$ *with* $n_i=\min(y_i,L-1)$. *Then, under* (4.5) *with* $\varepsilon_{il}\sim \mathrm{N}(0,1)$ *independently for* i=$1,\ldots,n$ *and* l=$1,\ldots,L-1$, *we have*

$$p(\mathbf{y}\mid\boldsymbol{\beta},\mathbf{X})=\prod_{i=1}^{n}p(y_i\mid\boldsymbol{\beta},\mathbf{x}_i)=\prod_{i=1}^{n}\Phi_{n_i}(\mathbf{X}_i\boldsymbol{\beta};\mathbf{I}_{n_i})=\Phi_{\bar{n}}(\bar{\mathbf{X}}\boldsymbol{\beta};\boldsymbol{\Lambda}), \tag{4.6}$$

where $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^\intercal,\ldots,\boldsymbol{\beta}_{L-1}^\intercal)^\intercal$, $\boldsymbol{\Lambda}=\mathbf{I}_{\bar{n}}$ *and* $\bar{\mathbf{X}}$ *is a* $\bar{n}\times(L-1)\cdot p$ *matrix with* $n_i\times(L-1)\cdot p$ *blocks* $\bar{\mathbf{X}}_{[i1]}=\mathbf{X}_i$, $i=1,\ldots,n$, *where* $\mathbf{X}_i=[\mathrm{diag}(2\bar{\mathbf{y}}_i-\mathbf{1})\otimes\mathbf{x}_i^\intercal,\mathbf{0}_{n_i\times(L-1-n_i)\cdot p}]$, *for* $i=1,\ldots,n$. *In* (4.6), *the generic quantity* $\mathbf{I}_n$ *refers to the* $n\times n$ *identity matrix.*

To clarify Proposition 4.3, it suffices to re-write $p(y_i=l\mid\boldsymbol{\beta},\mathbf{x}_i)$, $l=1,\ldots,L-1$, in (4.5), as $\Phi(\mathbf{x}_i^\intercal\boldsymbol{\beta}_l)\prod_{k=1}^{l-1}[1-\Phi(\mathbf{x}_i^\intercal\boldsymbol{\beta}_k)]=\prod_{k=1}^{l}\Phi[(2\bar{\mathbf{y}}_{ik}-1)\mathbf{x}_i^\intercal\boldsymbol{\beta}_k]=\Phi_l(\mathbf{X}_i\boldsymbol{\beta};\mathbf{I}_l)$, where $\bar{\mathbf{y}}_i$ is defined as in Proposition 4.3. The above result leverages standard properties of multivariate Gaussians.

Combining Propositions 4.1–4.3 it is clear that, despite characterizing different utility–based decision mechanisms, models (4.1), (4.3) and (4.5) have the same form for the joint likelihood of the observed responses. The only difference among such likelihoods is their dimension and the definition of the known matrices $\bar{\mathbf{X}}$ and $\boldsymbol{\Lambda}$, which change depending on the type of model. These results are fundamental for the novel conjugacy results in Section 4.3.

4.3 Bayesian Inference for the Multinomial Probit Models

Common Bayesian implementations of multinomial probit models consider a multivariate Gaussian prior $\mathrm{N}_q(\boldsymbol{\xi},\boldsymbol{\Omega})$ for the parameters in $\boldsymbol{\beta}$, where q is equal to p in model (4.1) and to $p\cdot(L-1)$ in models (4.3) and (4.5), whereas $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$ denote the pre–specified prior mean vector and covariance matrix, respectively (Albert and Chib, 1993; McCulloch and Rossi, 1994; Nobile, 1998; McCulloch et al., 2000; Albert and Chib, 2001; Chen and Kuo, 2002; Imai and Van Dyk, 2005; Zhang et al., 2006; Burgette and Nordheim, 2012; Johndrow et al., 2013). Besides providing a default specification in various Bayesian regression models, this choice is also motivated by the Gaussian assumption for the latent utilities in

models (4.1), (4.3) and (4.5) which implies an augmented–data representation facilitating the implementation of MCMC (Albert and Chib, 1993, 2001; Imai and Van Dyk, 2005; Holmes and Held, 2006; Chopin and Ridgway, 2017) and approximate methods (Girolami and Rogers, 2006; Girolami and Zhong, 2007; Riihimäki et al., 2013; Knowles and Minka, 2011) for inference and prediction.

As discussed in Section 4.1, the above strategies have computational drawbacks—especially in large p settings—and are motivated by the apparent absence of conjugacy between multinomial probit likelihoods and Gaussian priors for the $\boldsymbol{\beta}$ parameters. In Section 4.3.1, we show not only that the posterior in this setting is a SUN, but also that the whole SUN family is conjugate to multinomial probits, thereby obtaining closed–form posterior distributions under a broad variety of priors, which include also the default Gaussian one and, as a byproduct, Gaussian processes. Leveraging the novel results in Section 4.3.1, we develop in Section 4.3.2 improved Monte Carlo methods for full Bayesian inference and classification, along with scalable and accurate approximations of the posterior in high–dimensional settings.

Before providing an overview of the SUN distribution (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014) and presenting our conjugacy results, we shall emphasize that some of the aforementioned contributions consider also priors for $\boldsymbol{\Sigma}$ in models (4.1) and (4.3). Our focus in this chapter is on the posterior for $\boldsymbol{\beta}$ conditioned on $\boldsymbol{\Sigma}$ and, hence, we avoid additional identifiability and computational complications which arise when including a prior also for $\boldsymbol{\Sigma}$. Nonetheless, as discussed in Section 4.5, the closed–form expression for the marginal likelihood $p(\mathbf{y} \mid \mathbf{X})$ presented in Corollary 4.5, and the i.i.d. sampler to generate values from the posterior $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})$ outlined in Algorithm 6, can be useful to estimate $\boldsymbol{\Sigma}$ via empirical Bayes, and to improve sampling of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ from their full–conditionals.

4.3.1 Conjugacy via unified skew–normal priors

Consistent with Section 4.3, and recalling Section 3.2, let us assume a $\text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ prior for $\boldsymbol{\beta}$, whose density follows from (3.6) and, for ease of reading, has the form

$$p(\boldsymbol{\beta}) = \phi_q(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \frac{\Phi_h(\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta})}{\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})}.$$

See Section 3.2 and references therein for a more detailed overview about the SUN distribution. Recall that, as specified in Section 3.2, when all the entries in $\boldsymbol{\Delta}$ are 0, $p(\boldsymbol{\beta})$ coincides with the density of a q–variate Gaussian with mean vector $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}$ obtained via the quadratic combination among the correlation matrix $\bar{\boldsymbol{\Omega}}$ and the diagonal scale matrix $\boldsymbol{\omega} = (\boldsymbol{\Omega} \odot \mathbf{I}_q)^{1/2}$, where $\odot$ is the element–wise Hadamard product. Such a class of Gaussian priors can be also obtained by setting $h = 0$. As discussed in Arellano-Valle and Azzalini (2006), the multivariate Gaussian case is just an example of

a broad variety of distributions which can be obtained from a $\text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ prior under suitable choices for its parameters. Priors of potential interest within this class are independent univariate skew–normals (Azzalini, 1985) for the parameters in $\boldsymbol{\beta}$ and classical multivariate skew–normals (Azzalini and Dalla Valle, 1996) for the entire vector $\boldsymbol{\beta}$. Therefore, our results allow tractable inference in Bayesian multinomial probit models under a broad class of priors that include Gaussian specifications along with asymmetric priors that may be useful in social science and econometric studies. Note that also non–linear formulations via Gaussian processes induce multivariate Gaussian priors and, hence, our results can be directly applied to the flexible classification strategies discussed in Girolami and Rogers (2006) and Riihimäki et al. (2013).

The main roles of the parameters $\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ is further clarified by the additive representation (3.7), which, if $\boldsymbol{\beta} \sim \text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, writes

$$\boldsymbol{\beta} \stackrel{d}{=} \boldsymbol{\xi} + \boldsymbol{\omega}(\mathbf{U}_0 + \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\mathbf{U}_1), \quad \mathbf{U}_0 \perp \mathbf{U}_1,$$

with $\mathbf{U}_0 \sim \text{N}_q(\mathbf{0}, \bar{\boldsymbol{\Omega}} - \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}^\intercal)$ and $\mathbf{U}_1$ from a $\text{N}_h(\mathbf{0}, \boldsymbol{\Gamma})$ truncated below $-\boldsymbol{\gamma}$, shortly denoted as $\text{TN}_h(-\boldsymbol{\gamma}; \mathbf{0}, \boldsymbol{\Gamma})$.

Besides clarifying the role of the prior parameters, this stochastic additive representation of the SUN random variable is useful also for posterior inference since, as already shown in Chapters 2 and 3 and as we will discuss in the following, it provides a direct strategy to sample i.i.d. values from the SUN distribution, thus improving upon state–of–the–art MCMC methods for Bayesian multinomial probit models. Indeed, as shown in Theorem 4.4, the SUN prior $\text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ is conjugate to the multinomial probit likelihoods reported in (4.2), (4.4) and (4.6), meaning that also the posterior $(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})$ has a SUN distribution. In particular we show that $(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) \sim \text{SUN}_{q,h+m}(\boldsymbol{\xi}_{\text{post}}, \boldsymbol{\Omega}_{\text{post}}, \boldsymbol{\Delta}_{\text{post}}, \boldsymbol{\gamma}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}})$.

Theorem 4.4. *Let $p(\boldsymbol{\beta})$ be a $\text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ prior density function and denote with $\Phi_m(\bar{\mathbf{X}}\boldsymbol{\beta}; \boldsymbol{\Lambda})$ the generic multinomial probit likelihood in equations (4.2), (4.4) and (4.6), with m, $\bar{\mathbf{X}}$ and $\boldsymbol{\Lambda}$ defined as in Propositions 4.1, 4.2 or 4.3 depending on whether the focus is on model (4.1), (4.3) or (4.5), respectively. Then, the posterior density $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})$ of $\boldsymbol{\beta}$ is*

$$p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \phi_q(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}; \boldsymbol{\Omega}_{\text{post}}) \frac{\Phi_{h+m}(\boldsymbol{\gamma}_{\text{post}} + \boldsymbol{\Delta}_{\text{post}}^\intercal \bar{\boldsymbol{\Omega}}_{\text{post}}^{-1} \boldsymbol{\omega}_{\text{post}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}); \boldsymbol{\Gamma}_{\text{post}} - \boldsymbol{\Delta}_{\text{post}}^\intercal \bar{\boldsymbol{\Omega}}_{\text{post}}^{-1} \boldsymbol{\Delta}_{\text{post}})}{\Phi_{h+m}(\boldsymbol{\gamma}_{\text{post}}; \boldsymbol{\Gamma}_{\text{post}})}, \tag{4.7}$$

with $\boldsymbol{\xi}_{\text{post}} = \boldsymbol{\xi}$, $\boldsymbol{\Omega}_{\text{post}} = \boldsymbol{\Omega}$, $\boldsymbol{\Delta}_{\text{post}} = (\boldsymbol{\Delta}, \bar{\boldsymbol{\Omega}}\boldsymbol{\omega}\bar{\mathbf{X}}^\intercal \mathbf{s}^{-1})$, $\boldsymbol{\gamma}_{\text{post}} = (\boldsymbol{\gamma}^\intercal, \boldsymbol{\xi}^\intercal \bar{\mathbf{X}}^\intercal \mathbf{s}^{-1})^\intercal$ and $\boldsymbol{\Gamma}_{\text{post}}$ is an $(h+m) \times (h+m)$ covariance matrix with blocks $\boldsymbol{\Gamma}_{\text{post}[11]} = \boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}_{\text{post}[22]} = \mathbf{s}^{-1}(\bar{\mathbf{X}}\boldsymbol{\Omega}\bar{\mathbf{X}}^\intercal + \boldsymbol{\Lambda})\mathbf{s}^{-1}$ and $\boldsymbol{\Gamma}_{\text{post}[21]} = \boldsymbol{\Gamma}^\intercal{}_{\text{post}[12]} = \mathbf{s}^{-1}\bar{\mathbf{X}}\boldsymbol{\omega}\boldsymbol{\Delta}$, where $\mathbf{s} = [(\bar{\mathbf{X}}\boldsymbol{\Omega}\bar{\mathbf{X}}^\intercal + \boldsymbol{\Lambda}) \odot \mathbf{I}_m]^{1/2}$. Note that in (4.7), the dimension q is equal to p under model (4.1) and to $p \cdot (L-1)$ under models (4.3) and (4.5).

As a consequence of Theorem 4.4, it follows that also the common multivariate Gaussian prior for $\boldsymbol{\beta}$—which is a special case of unified skew–normal—leads to a SUN posterior

when updated with the multinomial probit likelihoods in (4.2), (4.4) and (4.6). In particular, if $p(\boldsymbol{\beta}) = \phi_q(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega})$ it immediately follows from Theorem 4.4 that the posterior distribution is a SUN having density as in (4.7), with $h = 0$ and posterior parameters $\boldsymbol{\xi}_{\text{post}} = \boldsymbol{\xi}$, $\boldsymbol{\Omega}_{\text{post}} = \boldsymbol{\Omega}$, $\boldsymbol{\Delta}_{\text{post}} = \bar{\boldsymbol{\Omega}}\boldsymbol{\omega}\bar{\mathbf{X}}^\intercal\mathbf{s}^{-1}$, $\boldsymbol{\gamma}_{\text{post}} = \mathbf{s}^{-1}\bar{\mathbf{X}}\boldsymbol{\xi}$, $\boldsymbol{\Gamma}_{\text{post}} = \mathbf{s}^{-1}(\bar{\mathbf{X}}\boldsymbol{\Omega}\bar{\mathbf{X}}^\intercal + \boldsymbol{\Lambda})\mathbf{s}^{-1}$, where $\mathbf{s} = [(\bar{\mathbf{X}}\boldsymbol{\Omega}\bar{\mathbf{X}}^\intercal + \boldsymbol{\Lambda}) \odot \mathbf{I}_m]^{1/2}$.

Theorem 4.4 provides novel results with important implications in Bayesian inference for multinomial probit models. As discussed by Arellano-Valle and Azzalini (2006) and Azzalini and Capitanio (2014), SUN distributions share several common properties with multivariate Gaussians. A key one is that such a family is closed under marginalization, linear combinations and conditioning. Within our context, this means that the marginal posteriors for each coefficient and linear combinations of interest—such as those defining the latent utilities—are still SUN and their parameters can be obtained via simple transformations of those in Theorem 4.4 (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014). According to (4.7), also the normalizing constant of the posterior is available in closed form and coincides with the cumulative distribution function $\Phi_{h+m}(\boldsymbol{\gamma}_{\text{post}}; \boldsymbol{\Gamma}_{\text{post}})$ of a multivariate Gaussian with $\mathbf{0}$ mean and covariance matrix $\boldsymbol{\Gamma}_{\text{post}}$, evaluated at $\boldsymbol{\gamma}_{\text{post}}$. As highlighted in Corollaries 4.5 and 4.6, this result is fundamental to obtain closed–form expressions of marginal likelihoods and predictive distributions that are useful for model selection and classification.

Corollary 4.5. *Under the settings of Theorem 4.4, the marginal likelihood can be expressed as*

$$p(\mathbf{y} \mid \mathbf{X}) = \frac{p(\mathbf{y}, \boldsymbol{\beta} \mid \mathbf{X})}{p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})} = \frac{\Phi_{h+m}(\boldsymbol{\gamma}_{\text{post}}; \boldsymbol{\Gamma}_{\text{post}})}{\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})}, \tag{4.8}$$

with $\boldsymbol{\gamma}_{\text{post}}$ and $\boldsymbol{\Gamma}_{\text{post}}$ defined as in Theorem 4.4.

Corollary 4.6. *Consider the expanded dataset in which, besides the original data $\mathbf{y}$ and $\mathbf{X}$, we also have an additional unit with predictors $\mathbf{x}_{\text{new}}$ and response $y_{\text{new}} = l$. Moreover, let m_l, $\bar{\mathbf{X}}_l$ and $\boldsymbol{\Lambda}_l$ be defined as in Propositions 4.1, 4.2 or 4.3 depending on whether the focus is on likelihoods (4.2), (4.4) or (4.6), respectively, for the expanded data. Then, under the settings of Theorem 4.4, we have that*

$$\text{pr}(y_{\text{new}} = l \mid \mathbf{y}, \mathbf{X}, \mathbf{x}_{\text{new}}) = \frac{p(y_{\text{new}} = l, \mathbf{y} \mid \mathbf{X}, \mathbf{x}_{\text{new}})}{p(\mathbf{y} \mid \mathbf{X}, \mathbf{x}_{\text{new}})} = \frac{\Phi_{h+m_l}(\boldsymbol{\gamma}_{l\text{post}}; \boldsymbol{\Gamma}_{l\text{post}})}{\Phi_{h+m}(\boldsymbol{\gamma}_{\text{post}}; \boldsymbol{\Gamma}_{\text{post}})}, \tag{4.9}$$

for each $l = 1, \ldots, L$, with $\boldsymbol{\gamma}_{\text{post}}$ and $\boldsymbol{\Gamma}_{\text{post}}$ as in Theorem 4.4, while $\boldsymbol{\gamma}_{l\text{post}}$ and $\boldsymbol{\Gamma}_{l\text{post}}$ coincide with $\boldsymbol{\gamma}_{\text{post}}$ and $\boldsymbol{\Gamma}_{\text{post}}$, evaluated at $\bar{\mathbf{X}}_l$ and $\boldsymbol{\Lambda}_l$.

Corollaries 4.5 and 4.6 facilitate closed–form Bayesian hypothesis testing, variable selection and classification without the need to rely on MCMC. Exploiting the moment generating functions of the SUN in Section 2.3 of Arellano-Valle and Azzalini (2006) and

the additional derivations in Azzalini and Bacchieri (2010); Gupta et al. (2013); Azzalini and Capitanio (2014); Durante (2019), closed–form expressions can be derived also for the posterior mean of $\boldsymbol{\beta}$, its covariance matrix and the cumulative distribution function, thus facilitating Bayesian estimation, uncertainty quantification and classification. These closed–form expressions require, however, evaluation of high–dimensional cumulative distribution functions and tedious derivations that do not facilitate calculation of more complex functionals, thus motivating the alternative computational methods presented in Section 4.3.2.

4.3.2 Computational methods

This section provides new computational methods for Bayesian multinomial probit models that exploit results in Section 4.3.1 to improve upon state–of–the–art solutions, especially in large q settings. In particular, in Section 4.3.2.1 we focus on Monte Carlo methods that, unlike current MCMC solutions, rely on independent and identically distributed samples from the exact SUN posterior. This strategy requires, however, to sample from $(h+m)$–variate truncated normals with full covariance matrix and, hence, becomes impractical as the sample size grows. To address this issue, we also propose in Section 4.3.2.2 a blocked partially–factorized variational Bayes that relaxes various independence assumptions of classical mean–field families to obtain improved and computationally efficient approximations, that almost perfectly match the exact posterior in large q settings, especially when $q > h + m$.

4.3.2.1 Monte Carlo methods via independent samples from the posterior

Complex functionals of the posterior can be effectively evaluated via Monte Carlo methods leveraging the stochastic representation of the SUN reported in equation (3.7) and recalled in Section 4.3.1. This equivalent construction allows to sample independent and identically distributed values from the SUN posterior in Theorem 4.4, via linear combinations among samples from multivariate Gaussians and multivariate truncated normals. Such a routine is described in Algorithm 6 and crucially avoids MCMC strategies, thus circumventing convergence and mixing issues commonly seen in Bayesian multinomial probit models (Johndrow et al., 2013), while allowing for parallel computing. One possible computational drawback of Algorithm 6 is the need to sample from multivariate truncated normals. Recent advances relying on minimax tilting methods (Botev, 2017) have made this task possible and computationally feasible for multivariate truncated normals with a dimension of few hundreds, thus making Algorithm 6 a computationally efficient strategy in small–to–moderate $h + m$ and large, potentially huge, q studies. As discussed by Chopin and Ridgway (2017), these large q settings are actually those where state–of–the–art MCMC methods, including efficient `STAN` implementations of the Hamiltonian

Algorithm 6: Strategy to sample from the SUN posterior in Theorem 4.4

for *t=1,...,T* **do**

- **[1]** Sample $\mathbf{U}_0^{(t)} \sim \mathrm{N}_q(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{\text{post}} - \boldsymbol{\Delta}_{\text{post}}\boldsymbol{\Gamma}_{\text{post}}^{-1}\boldsymbol{\Delta}_{\text{post}}^{\intercal})$ [in `R` use the function `rmvnorm`]
- **[2]** Sample $\mathbf{U}_1^{(t)} \sim \mathrm{TN}_{h+m}(-\boldsymbol{\gamma}_{\text{post}}; \mathbf{0}, \boldsymbol{\Gamma}_{\text{post}})$ [in `R` use the function `rtmvnorm`]
- **[3]** Compute $\boldsymbol{\beta}^{(t)} = \boldsymbol{\xi}_{\text{post}} + \boldsymbol{\omega}_{\text{post}}(\mathbf{U}_0^{(t)} + \boldsymbol{\Delta}_{\text{post}}\boldsymbol{\Gamma}_{\text{post}}^{-1}\mathbf{U}_1^{(t)})$

Output: Independent and identically distributed samples $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(T)}$ from the posterior in Theorem 4.4. Based on such samples, posterior functionals $\mathbb{E}[g(\boldsymbol{\beta}) \mid \mathbf{y}, \mathbf{X}]$ can be computed, via Monte Carlo, as $\sum_{t=1}^{T} g(\boldsymbol{\beta}^{(t)})/T$.

no–u–turn sampler (Hoffman and Gelman, 2014), are computationally unfeasible. The results in Botev (2017) are also useful to evaluate efficiently cumulative distribution functions of multivariate Gaussians, and hence are practically relevant to calculate marginal likelihoods (4.8) and predictive probabilities (4.9) in small–to–moderate $h + m$ settings.

4.3.2.2 Blocked partially–factorized variational Bayes

As discussed in Section 4.3.2.1, when $h + m$ is large, sampling from $(h + m)$–variate truncated normals with full covariance matrix becomes computationally unfeasible (Botev, 2017), thus making Algorithm 6 impractical in such settings. Typically, h is either 0—when Gaussian priors are considered—or a small value, whereas m depends on the sample size n and on the number of classes L. Hence, it is necessary to devise more scalable methods, especially in common settings where n is larger than a few hundreds.

A possible solution to the above problem is to consider approximations of the posterior density, with variational Bayes providing a well–established procedure, especially in those models admitting simple augmented data representations (Blei et al., 2017). As clarified in Section 4.2, this is the case of multinomial probit models relying on Gaussian latent utilities. Such a property has motivated several variational strategies to approximate the joint posterior $p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})$ of $\boldsymbol{\beta}$ and the augmented data $\bar{\mathbf{z}}$, with a tractable density $q^*(\boldsymbol{\beta}, \bar{\mathbf{z}})$, which is the closest in Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) to $p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})$—among all the densities in a pre–specified approximating family $\mathcal{Q}$. As in the development of simple Gibbs samplers based on tractable full–conditionals (Albert and Chib, 1993), the inclusion of the augmented data facilitates the implementation of simple coordinate ascent variational inference (CAVI) routines (Bishop, 2006; Blei et al., 2017) to minimize, with respect to $q(\boldsymbol{\beta}, \bar{\mathbf{z}})$, the divergence $\mathrm{KL}[q(\boldsymbol{\beta}, \bar{\mathbf{z}}) || p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})]$.

Clearly, the availability of simple optimization routines and strategies to derive the optimal marginal $q^*(\boldsymbol{\beta})$ from $q^*(\boldsymbol{\beta}, \bar{\mathbf{z}})$, depend also on how the family $\mathcal{Q}$ is defined. Common solutions in binary (Consonni and Marin, 2007) and multinomial (Girolami and Rogers, 2006) probit settings rely on mean–field families $\mathcal{Q}_{\text{MF}} = \{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) : q(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q(\boldsymbol{\beta})q(\bar{\mathbf{z}})\}$ that assume independence between $\boldsymbol{\beta}$ and $\bar{\mathbf{z}}$. These strategies come with simple CAVI al-

gorithms which scale easily to high–dimensional settings and, due to the factorized form of $q^*(\boldsymbol{\beta}, \bar{\mathbf{z}})$, provide directly the approximating density $q^*(\boldsymbol{\beta})$ of interest. However, theoretical and empirical studies on simple univariate binary probit models considered in Chapter 2 show that this mean–field assumption often leads to low–quality approximations in high–dimensional probit settings, that massively affect not only uncertainty quantification, but also estimation and classification. To address this issue in the context of univariate binary probit models with Gaussian priors, in Chapter 2 we considered a partially factorized mean–field approximating family $\mathcal{Q}_{\text{PFM}} = \{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) : q(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) \prod_{r=1}^{h+m} q(\bar{z}_r)\}$ which avoids enforcing independence between $\boldsymbol{\beta}$ and $\bar{\mathbf{z}}$, and only assumes that $q(\bar{\mathbf{z}})$ factorizes as the product of its marginals. Such a new class of approximating densities substantially improves the quality of the original mean–field approximation and almost perfectly matches the exact posterior in high–dimensional settings, especially when the number of predictors is higher than the sample size, without sacrificing computational tractability.

Motivated by the above discussion, we develop a new blocked partially–factorized mean–field approximation that extends the contribution of Chapter 2 in three main important directions. In particular, we [i] allow the inclusion of SUN and not only Gaussian priors, [ii] generalize the methods to multinomial probit models, and [iii] further enlarge the class of approximating densities by replacing $\prod_{r=1}^{h+m} q(\bar{z}_r)$ in $\mathcal{Q}_{\text{PFM}}$ with $\prod_{c=1}^{C} q(\bar{\mathbf{z}}_c)$, where $\bar{\mathbf{z}}_1, \ldots, \bar{\mathbf{z}}_C$ are distinct sub–vectors of $\bar{\mathbf{z}}$, such that $\bar{\mathbf{z}} = (\bar{\mathbf{z}}_1^\intercal, \ldots, \bar{\mathbf{z}}_C^\intercal)^\intercal$. Hence, instead of enforcing independence among all the augmented data, we only make this assumption between pre–specified blocks. In fact, while in high–dimensional univariate binary settings the independence among all the augmented data does not seem to have a major impact on the quality of the approximation (see Chapter 2), this may not be the case in multinomial probit models. For example, under the formulation presented in Section 4.2.2, each unit i enters the matrix $\bar{\mathbf{X}}$ multiple times and, hence, it is reasonable to expect a relatively strong dependence among unit–specific augmented data, which cannot be accurately approximated by a fully factorized representation for $q(\bar{\mathbf{z}})$. Similar blocking ideas have been considered by Chopin (2011); Genton et al. (2018) and Cao et al. (2019), to simulate from multivariate truncated normals and compute cumulative distribution functions of high–dimensional Gaussians. We adapt these ideas in the context of variational inference to obtain improved approximations of the posterior, without affecting computational performance.

To introduce the blocked partially–factorized mean–field approximation, first note that the kernel of the posterior density $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})$ in (4.7) can be re–written as

$$p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) \propto \phi_q(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}; \boldsymbol{\Omega}_{\text{post}}) \int \phi_{h+m}[\bar{\mathbf{z}} - (\boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\beta}); \boldsymbol{\Sigma}_{\text{post}}] \mathbb{1}(\bar{\mathbf{z}} > \mathbf{0}) \mathrm{d}\bar{\mathbf{z}}, \quad (4.10)$$

with $\mathbf{X}_{\text{post}} = \boldsymbol{\Delta}_{\text{post}}^\intercal \bar{\boldsymbol{\Omega}}_{\text{post}}^{-1} \boldsymbol{\omega}_{\text{post}}^{-1}$, $\boldsymbol{\eta}_{\text{post}} = \boldsymbol{\gamma}_{\text{post}} - \mathbf{X}_{\text{post}} \boldsymbol{\xi}_{\text{post}}$, and $\boldsymbol{\Sigma}_{\text{post}} = \boldsymbol{\Gamma}_{\text{post}} - \boldsymbol{\Delta}_{\text{post}}^\intercal \bar{\boldsymbol{\Omega}}_{\text{post}}^{-1} \boldsymbol{\Delta}_{\text{post}}$. To clarify the connection between expression (4.7) and (4.10) it is sufficient to notice that

the integral in (4.10) coincides with $p(\bar{\mathbf{z}} > 0)$, when $\bar{\mathbf{z}} \sim \mathrm{N}_{h+m}(\boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\text{post}})$. In particular, $p(\bar{\mathbf{z}} > 0) = p[-(\bar{\mathbf{z}} - \boldsymbol{\eta}_{\text{post}} - \mathbf{X}_{\text{post}}\boldsymbol{\beta}) < \boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\beta}] = \Phi_{h+m}(\boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\beta}; \boldsymbol{\Sigma}_{\text{post}})$, which coincides with the cumulative distribution function in the numerator of equation (4.7). Leveraging such an alternative representation and Gaussian–Gaussian conjugacy, we can easily notice that

$$\begin{aligned} p(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \mathbf{y}, \mathbf{X}) &\propto \phi_q(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}; \boldsymbol{\Omega}_{\text{post}})\phi_{h+m}[\bar{\mathbf{z}} - (\boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\beta}); \boldsymbol{\Sigma}_{\text{post}}] \\ &\propto \phi_q(\boldsymbol{\beta} - \mathbf{V}_{\text{post}}[\mathbf{X}_{\text{post}}^\intercal \boldsymbol{\Sigma}_{\text{post}}^{-1}(\bar{\mathbf{z}} - \boldsymbol{\eta}_{\text{post}}) + \boldsymbol{\Omega}_{\text{post}}^{-1}\boldsymbol{\xi}_{\text{post}}]; \mathbf{V}_{\text{post}}), \end{aligned} \tag{4.11}$$

with $\mathbf{V}_{\text{post}} = (\mathbf{X}_{\text{post}}^\intercal \boldsymbol{\Sigma}_{\text{post}}^{-1}\mathbf{X}_{\text{post}} + \boldsymbol{\Omega}_{\text{post}}^{-1})^{-1}$. Thus,

$$(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \mathbf{y}, \mathbf{X}) \sim \mathrm{N}_q(\mathbf{V}_{\text{post}}[\mathbf{X}_{\text{post}}^\intercal \boldsymbol{\Sigma}_{\text{post}}^{-1}(\bar{\mathbf{z}} - \boldsymbol{\eta}_{\text{post}}) + \boldsymbol{\Omega}_{\text{post}}^{-1}\boldsymbol{\xi}_{\text{post}}], \mathbf{V}_{\text{post}}).$$

On the other hand, according to (4.10), the conditional density $p(\bar{\mathbf{z}} \mid \boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ of the augmented data $\bar{\mathbf{z}}$ is a multivariate normal with mean $\boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\beta}$, covariance matrix $\boldsymbol{\Sigma}_{\text{post}}$ and truncation below $\mathbf{0}$. Hence, by marginalizing out $\boldsymbol{\beta}$ with density $\phi_q(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{post}}; \boldsymbol{\Omega}_{\text{post}})$, we obtain

$$\begin{aligned} p(\bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X}) &\propto \phi_{h+m}[\bar{\mathbf{z}} - (\boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\xi}_{\text{post}}); \boldsymbol{\Sigma}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\Omega}_{\text{post}}\mathbf{X}_{\text{post}}^\intercal]\mathbb{1}(\bar{\mathbf{z}} > \mathbf{0}) \\ &\propto \phi_{h+m}[\bar{\mathbf{z}} - (\boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\xi}_{\text{post}}); \boldsymbol{\Gamma}_{\text{post}}]\mathbb{1}(\bar{\mathbf{z}} > \mathbf{0}), \end{aligned} \tag{4.12}$$

where $\boldsymbol{\Sigma}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\Omega}_{\text{post}}\mathbf{X}_{\text{post}}^\intercal = \boldsymbol{\Gamma}_{\text{post}}$. Combining (4.11)–(4.12), and recalling our discussion on variational Bayes, we aim to provide an accurate approximation $q^*(\boldsymbol{\beta}, \bar{\mathbf{z}})$ of the joint density

$$\begin{aligned} p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X}) &= p(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \mathbf{y}, \mathbf{X})p(\bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X}) \\ &\propto \phi_q(\boldsymbol{\beta} - \mathbf{V}_{\text{post}}[\mathbf{X}_{\text{post}}^\intercal \boldsymbol{\Sigma}_{\text{post}}^{-1}(\bar{\mathbf{z}} - \boldsymbol{\eta}_{\text{post}}) + \boldsymbol{\Omega}_{\text{post}}^{-1}\boldsymbol{\xi}_{\text{post}}]; \mathbf{V}_{\text{post}}) \\ &\quad \times \phi_{h+m}[\bar{\mathbf{z}} - (\boldsymbol{\eta}_{\text{post}} + \mathbf{X}_{\text{post}}\boldsymbol{\xi}_{\text{post}}); \boldsymbol{\Gamma}_{\text{post}}]\mathbb{1}(\bar{\mathbf{z}} > \mathbf{0}), \end{aligned} \tag{4.13}$$

such that $q^*(\boldsymbol{\beta}, \bar{\mathbf{z}})$ minimizes the KL divergence $\mathrm{KL}[q(\boldsymbol{\beta}, \bar{\mathbf{z}}) || p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})]$ within the blocked partially–factorized mean–field family of distributions $\mathfrak{Q}_{\text{PFM-B}} = \{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) : q(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) \prod_{c=1}^{C} q(\bar{\mathbf{z}}_c)\}$, where $\bar{\mathbf{z}}_1, \ldots, \bar{\mathbf{z}}_C$ are pre–specified sub–vectors of $\bar{\mathbf{z}}$. Equation (4.13) clarifies why $\mathfrak{Q}_{\text{PFM-B}}$ provides a particularly suitable family of approximating densities for $p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})$. In particular, since the exact conditional density $p(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \mathbf{y}, \mathbf{X})$ has a tractable Gaussian form, assuming independence between $\boldsymbol{\beta}$ and $\bar{\mathbf{z}}$ as in classical mean–field variational Bayes seems an unnecessarily strong assumption in this case. On the other hand, the main source of intractability in $p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})$ arises from the high–dimensional truncated normal density $p(\bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})$ with full covariance matrix $\boldsymbol{\Gamma}_{\text{post}}$, thus motivating our attempt to approximate it via a set of C independent lower–dimensional truncated normal densities $q^*(\bar{\mathbf{z}}_1) \cdots q^*(\bar{\mathbf{z}}_C)$. Each of these blocks must be sufficiently small to allow tractable inference under the associated truncated normal approximation, and should be specified so as to group augmented data with strong correlations in $\boldsymbol{\Gamma}_{\text{post}}$. Remark 4.7 discusses and motivates a possible default strategy to define the different blocks in multinomial probit models, when necessary.

Remark 4.7. *In multinomial probit models, when necessary, it is typically sufficient to group the augmented data associated with the same unit i, provided that there may be a strong overlap in the rows of $\bar{\mathbf{X}}$ referring to i, thus leading to high correlation in $\boldsymbol{\Gamma}_{\text{post}}$. Such a choice is further motivated by the fact that optimal mean–field solutions $q^*_{\text{MF}}(\boldsymbol{\beta}, \bar{\mathbf{z}})$ — which do not assume a factorized form for $q(\bar{\mathbf{z}})$ in $\mathcal{Q}_{\text{MF}} = \{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) : q(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q(\boldsymbol{\beta})q(\bar{\mathbf{z}})\}$ — are defined as $q^*_{\text{MF}}(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q^*_{\text{MF}}(\boldsymbol{\beta}) \prod_{i=1}^n q^*_{\text{MF}}(\bar{\mathbf{z}}_i)$ (Girolami and Rogers, 2006). Such a solution belongs also to $\mathcal{Q}_{\text{PFM-B}}$ when blocking units i and, hence, $min_{q(\boldsymbol{\beta},\bar{\mathbf{z}})\in\mathcal{Q}_{\text{PFM-B}}} \text{KL}[q(\boldsymbol{\beta}, \bar{\mathbf{z}})||p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})] \leq min_{q(\boldsymbol{\beta},\bar{\mathbf{z}})\in\mathcal{Q}_{\text{MF}}} \text{KL}[q(\boldsymbol{\beta}, \bar{\mathbf{z}})||p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})]$. In addition, we have $min_{q(\boldsymbol{\beta},\bar{\mathbf{z}})\in\mathcal{Q}_{\text{PFM-B}}} \text{KL}[q(\boldsymbol{\beta}, \bar{\mathbf{z}})||p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})] \leq min_{q(\boldsymbol{\beta},\bar{\mathbf{z}})\in\mathcal{Q}_{\text{PFM}}} \text{KL}[q(\boldsymbol{\beta}, \bar{\mathbf{z}})||p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})]$ since $\mathcal{Q}_{\text{PFM}} \subset \mathcal{Q}_{\text{PFM-B}}$ for any blocking structure, . Therefore, when blocking across statistical units, our optimal solution is guaranteed to improve both classical mean–field variational Bayes and recent partially factorized solutions.*

Besides providing a wider and more flexible class, the family $\mathcal{Q}_{\text{PFM-B}}$ also allows straightforward optimization as shown in Proposition 4.8.

Proposition 4.8. *The* KL *divergence* $\text{KL}[q(\boldsymbol{\beta}, \bar{\mathbf{z}})||p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})]$ *between* $p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})$ *in* (4.13) *and* $q(\boldsymbol{\beta}, \bar{\mathbf{z}}) \in \mathcal{Q}_{\text{PFM-B}}$, *is minimized at* $q^*(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) \prod_{c=1}^C q^*(\bar{\mathbf{z}}_c)$, *with*

$$q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) \propto \phi_q(\boldsymbol{\beta} - \mathbf{V}_{\text{post}}[\mathbf{X}^\intercal_{\text{post}} \boldsymbol{\Sigma}^{-1}_{\text{post}}(\bar{\mathbf{z}} - \boldsymbol{\eta}_{\text{post}}) + \boldsymbol{\Omega}^{-1}_{\text{post}} \boldsymbol{\xi}_{\text{post}}]; \mathbf{V}_{\text{post}}), \tag{4.14}$$

$$q^*(\bar{\mathbf{z}}_c) \propto \phi_{n_c}[\bar{\mathbf{z}}_c - \boldsymbol{\eta}_{[c]\text{post}} - \mathbf{W}_{[c]\text{post}}(\mathbb{E}_{q^*(\bar{\mathbf{z}}_{-c})}(\bar{\mathbf{z}}_{-c}) - \boldsymbol{\eta}_{[-c]\text{post}}); \boldsymbol{\Gamma}_{[c]\text{post}}] \mathbb{1}(\bar{\mathbf{z}}_c > \mathbf{0}), \ \forall c, \tag{4.15}$$

where $\mathbf{W}_{[c]\text{post}} = \boldsymbol{\Gamma}_{[c,-c]\text{post}} \boldsymbol{\Gamma}^{-1}_{[-c,-c]\text{post}}$ *and* $\boldsymbol{\Gamma}_{[c]\text{post}} = \boldsymbol{\Gamma}_{[c,c]\text{post}} - \boldsymbol{\Gamma}_{[c,-c]\text{post}} \boldsymbol{\Gamma}^{-1}_{[-c,-c]\text{post}} \boldsymbol{\Gamma}_{[-c,c]\text{post}}$, *with* $\boldsymbol{\Gamma}_{[c,c]\text{post}}$, $\boldsymbol{\Gamma}_{[-c,-c]\text{post}}$, $\boldsymbol{\Gamma}_{[-c,c]\text{post}}$ *and* $\boldsymbol{\Gamma}_{[c,-c]\text{post}}$ *denoting blocks of* $\boldsymbol{\Gamma}_{\text{post}}$ *when partitioned to highlight the sub–vector* $\bar{\mathbf{z}}_c$ *against all the others in* $\bar{\mathbf{z}}_{-c}$. *Similarly,* $\boldsymbol{\eta}_{[c]\text{post}}$ *and* $\boldsymbol{\eta}_{[-c]\text{post}}$ *denote the sub–vectors of* $\boldsymbol{\eta}_{\text{post}}$ *corresponding to block* c *and to all the other blocks, respectively. Finally,*

$$\mathbb{E}_{q^*(\bar{\mathbf{z}}_{-c})}(\bar{\mathbf{z}}_{-c}) = [\mathbb{E}_{q^*(\bar{\mathbf{z}}_1)}(\bar{\mathbf{z}}_1)^\intercal, \ldots, \mathbb{E}_{q^*(\bar{\mathbf{z}}_{c-1})}(\bar{\mathbf{z}}_{c-1})^\intercal, \mathbb{E}_{q^*(\bar{\mathbf{z}}_{c+1})}(\bar{\mathbf{z}}_{c+1})^\intercal, \ldots, \mathbb{E}_{q^*(\bar{\mathbf{z}}_C)}(\bar{\mathbf{z}}_C)^\intercal]^\intercal,$$

where the expectations are taken with respect to the optimal truncated normal approximations.

The solution in equation (4.14) is a direct consequence of the chain rule for the KL divergence. In fact, $\text{KL}[q(\boldsymbol{\beta}, \bar{\mathbf{z}})||p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})] = \text{KL}[q(\bar{\mathbf{z}})||p(\bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})] + \mathbb{E}_{q(\bar{\mathbf{z}})}\{\text{KL}[q(\boldsymbol{\beta} \mid \bar{\mathbf{z}})||p(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \mathbf{y}, \mathbf{X})]\}$, and hence the non–negative second summand is exactly zero for every $q(\bar{\mathbf{z}})$ only when $q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) = p(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \mathbf{y}, \mathbf{X})$. To clarify equation (4.15) recall that the optimal solution for $q(\bar{\mathbf{z}}_c)$ is proportional to $\exp[\mathbb{E}_{q^*(\bar{\mathbf{z}}_{-c})}(\log[p(\bar{\mathbf{z}}_c \mid \bar{\mathbf{z}}_{-c}, \mathbf{y}, \mathbf{X})])]$ (Bishop, 2006; Blei et al., 2017). Hence, recalling Horrace (2005) and Holmes and Held (2006), since $(\bar{\mathbf{z}} \mid \mathbf{y}, \mathbf{X})$ has a multivariate truncated Gaussian density (4.12), it follows that also each $p(\bar{\mathbf{z}}_c \mid \bar{\mathbf{z}}_{-c}, \mathbf{y}, \mathbf{X})$ is an n_c–variate truncated normal density, whose log–kernel is linear

Algorithm 7: CAVI for blocked partially–factorized approximation in Proposition 4.8

for *t=1 until convergence* **do**

  for *c=1, ..., C* **do**

    Set $\mathbb{E}_{q^{(t)}(\bar{\mathbf{z}}_c)}(\bar{\mathbf{z}}_c)$ equal to the expected value of an n_c–variate Gaussian with mean $\boldsymbol{\eta}_{[c]\text{post}} + \mathbf{W}_{[c]\text{post}}(\mathbb{E}_{q^{(t-1)}(\bar{\mathbf{z}}_{-c})}(\bar{\mathbf{z}}_{-c}) - \boldsymbol{\eta}_{[-c]\text{post}})$, covariance matrix $\boldsymbol{\Gamma}_{[c]\text{post}}$ and truncation below $\mathbf{0}$, where the expectations $\mathbb{E}_{q^{(t-1)}(\bar{\mathbf{z}}_{-c})}(\bar{\mathbf{z}}_{-c})$ are defined as

$$[\mathbb{E}_{q^{(t)}(\bar{\mathbf{z}}_1)}(\bar{\mathbf{z}}_1)^\intercal, \ldots, \mathbb{E}_{q^{(t)}(\bar{\mathbf{z}}_{c-1})}(\bar{\mathbf{z}}_{c-1})^\intercal, \mathbb{E}_{q^{(t-1)}(\bar{\mathbf{z}}_{c+1})}(\bar{\mathbf{z}}_{c+1})^\intercal, \ldots, \mathbb{E}_{q^{(t-1)}(\bar{\mathbf{z}}_C)}(\bar{\mathbf{z}}_C)^\intercal]^\intercal.$$

    [in `R` use the function `MomTrunc` to compute the mean of truncated normals].

Output: Optimal truncated normal approximating densities $q^*(\bar{\mathbf{z}}_1), \ldots, q^*(\bar{\mathbf{z}}_C)$ from (4.15), which are then combined with the closed–form solution for $q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}})$ (4.14), to provide the optimal joint approximating density $q^*(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) \prod_{c=1}^C q^*(\bar{\mathbf{z}}_c)$.

in $\bar{\mathbf{z}}_{-c}$ and the remaining parameters are specified as in (4.15). As is clear from Proposition 4.8, the only unknown parameters are $\mathbb{E}_{q^*(\bar{\mathbf{z}}_c)}(\bar{\mathbf{z}}_c)$, $c = 1, \ldots, C$, whose solution requires solving a non–linear system of equations. Algorithm 7 summarizes the key steps of the CAVI to obtain such quantities via simple operations.

Once $q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}})$ and $q^*(\bar{\mathbf{z}}) = \prod_{c=1}^C q^*(\bar{\mathbf{z}}_c)$ are available, approximations of key functionals of $\boldsymbol{\beta}$ can be easily derived leveraging the law of total expectation and results in Proposition 4.8. In particular, since $\mathbb{E}_{q^*(\boldsymbol{\beta})}(\boldsymbol{\beta}) = \mathbb{E}_{q^*(\bar{\mathbf{z}})}[\mathbb{E}_{q^*(\boldsymbol{\beta}|\bar{\mathbf{z}})}(\boldsymbol{\beta})]$, we have that

$$\mathbb{E}_{q^*(\boldsymbol{\beta})}(\boldsymbol{\beta}) = \mathbf{V}_{\text{post}}[\mathbf{X}_{\text{post}}^\intercal \boldsymbol{\Sigma}_{\text{post}}^{-1}(\mathbb{E}_{q^*(\bar{\mathbf{z}})}(\bar{\mathbf{z}}) - \boldsymbol{\eta}_{\text{post}}) + \boldsymbol{\Omega}_{\text{post}}^{-1}\boldsymbol{\xi}_{\text{post}}], \tag{4.16}$$

whereas, the equality $\text{var}_{q^*(\boldsymbol{\beta})}(\boldsymbol{\beta}) = \mathbb{E}_{q^*(\bar{\mathbf{z}})}[\text{var}_{q^*(\boldsymbol{\beta}|\bar{\mathbf{z}})}(\boldsymbol{\beta})] + \text{var}_{q^*(\bar{\mathbf{z}})}[\mathbb{E}_{q^*(\boldsymbol{\beta}|\bar{\mathbf{z}})}(\boldsymbol{\beta})]$, leads to

$$\text{var}_{q^*(\boldsymbol{\beta})}(\boldsymbol{\beta}) = \mathbf{V}_{\text{post}} + \mathbf{V}_{\text{post}}\mathbf{X}_{\text{post}}^\intercal \boldsymbol{\Sigma}_{\text{post}}^{-1} \text{var}_{q^*(\bar{\mathbf{z}})}(\bar{\mathbf{z}}) \boldsymbol{\Sigma}_{\text{post}}^{-1} \mathbf{X}_{\text{post}} \mathbf{V}_{\text{post}}. \tag{4.17}$$

To evaluate (4.16) and (4.17), it is sufficient to compute $\mathbb{E}_{q^*(\bar{\mathbf{z}}_c)}(\bar{\mathbf{z}}_c)$ and $\text{var}_{q^*(\bar{\mathbf{z}}_c)}(\bar{\mathbf{z}}_c)$, separately for each $c = 1, \ldots, C$, since due to the independence assumption among the C sub–vectors of $\bar{\mathbf{z}}$, the vector $\mathbb{E}_{q^*(\bar{\mathbf{z}})}(\bar{\mathbf{z}})$ has blocks $\mathbb{E}_{q^*(\bar{\mathbf{z}})}(\bar{\mathbf{z}})_{[c]} = \mathbb{E}_{q^*(\bar{\mathbf{z}}_c)}(\bar{\mathbf{z}}_c)$ for each $c = 1, \ldots, C$, whereas $\text{var}_{q^*(\bar{\mathbf{z}})}(\bar{\mathbf{z}})$ is a block–diagonal matrix with $\text{var}_{q^*(\bar{\mathbf{z}})}(\bar{\mathbf{z}})_{[cc]} = \text{var}_{q^*(\bar{\mathbf{z}}_c)}(\bar{\mathbf{z}}_c)$. As mentioned previously, in multinomial probit models such blocks typically refer to rows in the design matrix $\bar{\mathbf{X}}$ corresponding to the same unit i and, hence, their dimensions $n_1, \ldots, n_C$ are, by definition, equal or lower than the number of classes L, which is small in most applications. This allows fast evaluation of $\mathbb{E}_{q^*(\bar{\mathbf{z}}_c)}(\bar{\mathbf{z}}_c)$ and $\text{var}_{q^*(\bar{\mathbf{z}}_c)}(\bar{\mathbf{z}}_c)$ via routine `R` functions, such as `MomTrunc`.

Although (4.16) and (4.17) are typically the main quantities of interest, other generic functionals $\mathbb{E}_{q^*(\boldsymbol{\beta})}[g(\boldsymbol{\beta})]$ can be easily derived via simple Monte Carlo methods based on samples from $q^*(\boldsymbol{\beta})$. Combining equation (4.14)–(4.15), such draws can be obtained by setting

$$\boldsymbol{\beta}^{(t)} = \mathbf{V}_{\text{post}}[\mathbf{X}_{\text{post}}^{\intercal}\boldsymbol{\Sigma}_{\text{post}}^{-1}([\bar{\mathbf{z}}_1^{(t)\intercal}, \dots, \bar{\mathbf{z}}_C^{(t)\intercal}]^{\intercal} - \boldsymbol{\eta}_{\text{post}}) + \boldsymbol{\Omega}_{\text{post}}^{-1}\boldsymbol{\xi}_{\text{post}}] + \boldsymbol{\varepsilon}^{(t)}, \tag{4.18}$$

for $t = 1, \dots, T$, where $\boldsymbol{\varepsilon}^{(t)} \sim \text{N}_q(\mathbf{0}, \mathbf{V}_{\text{post}})$, and $\bar{\mathbf{z}}_c^{(t)} \sim \text{TN}_{n_c}[\mathbf{0}; \boldsymbol{\eta}_{[c]\text{post}} + \mathbf{W}_{[c]\text{post}}(\mathbb{E}_{q^*(\bar{\mathbf{z}}_{-c})}(\bar{\mathbf{z}}_{-c}) - \boldsymbol{\eta}_{[-c]\text{post}}), \boldsymbol{\Gamma}_{[c]\text{post}}]$ for $c = 1, \dots, C$. Also in this case, since n_c is typically very small, samples from n_c–variate truncated normal can be effectively obtained from common `R` functions, such as `rtmvnorm`. Such a Monte Carlo strategy is particularly useful to compute the predictive probabilities for a new unit with covariates $\mathbf{x}_{\text{new}}$. To accomplish this goal, it is sufficient to compute, for each sample $\boldsymbol{\beta}^{(t)}$ of $\boldsymbol{\beta}$, the latent utilities $z_{\text{new}l}^{(t)}$, $l = 1, \dots, L$ defined either via (4.1), (4.3) or (4.5), depending on the type of multinomial probit model considered. Then, if the focus is on models (4.1) and (4.3), a Monte Carlo estimate for $p(y_{\text{new}} = l \mid \mathbf{y}, \mathbf{X}, \mathbf{x}_{\text{new}})$ can be obtained by computing the relative frequency of samples in which $z_{\text{new}l}^{(t)} > z_{\text{new}k}^{(t)}$ for all $k \neq l$. If, instead, one considers the sequential representation in (4.5), the Monte Carlo estimate for $p(y_{\text{new}} = l \mid \mathbf{y}, \mathbf{X}, \mathbf{x}_{\text{new}})$ coincides with the relative frequency of samples in which $z_{\text{new}l}^{(t)} > 0$ and $z_{\text{new}k}^{(t)} < 0$, for all $k < l$.

4.4 Gastrointestinal Lesions Application

To evaluate the performances of the methods developed in Section 4.3, we consider a medical study by Mesejo et al. (2016) which focuses on 76 gastrointestinal lesions classified as `hyperplasic` ($l = 1$), `serrated adenoma` ($l = 2$) and `adenoma` ($l = 3$), where the first is benign, while the others are malignant. For each lesion, a vector of 1396 features is available and comprises 2D textural, 2D color, and 3D shape data, measured with both white light and narrow band imaging. In our analyses we standardized the predictors as suggested by Gelman et al. (2008) and Chopin and Ridgway (2017), and removed features that were always 0, thus obtaining $p - 1 = 929$ predictors $\mathbf{x}_i \in \mathbb{R}^{p-1}$ with mean 0 and standard deviation 0.5. To assess predictive performance, we also held out 15 randomly chosen observations from the calculation of the posterior, roughly corresponding to 20% of the dataset.

As already discussed in Section 4.1, Bayesian inference for such an high–dimensional study may be computationally unfeasible under state–of–the–art MCMC methods (Chopin and Ridgway, 2017), and hence it provides a useful setting for quantifying to what extent the new methods developed in Sections 4.3 can cover such a gap. To do this, we first focus on the sequential discrete choice multinomial probit model in Section 4.2.3 with Gaussian priors, and compare the computational performance of the methods developed

in Section 4.3.2 with the `rstan` implementation of the Hamiltonian no–u–turn sampler in Hoffman and Gelman (2014). The choice of the sequential model is directly motivated by the type of response of interest in our study. Indeed, it is plausible to first model benign ($l = 1$) against malignant ($l > 1$) status, and then focus on comparing the two sub–categories $l = 2$ and $l = 3$ of malignant lesions. Under this model, the vector $\boldsymbol{\beta}$ has dimension 1860, corresponding to the two class–specific 929–dimensional parameter vectors plus a class–specific intercept term. Consistent with Albert and Chib (2001), we place a $\text{N}_{1860}(\mathbf{0}, \omega^2\mathbf{I}_{1860})$ prior on $\boldsymbol{\beta}$, with $\omega = 5$ in line with guidelines in Gelman et al. (2008).

Figure 4.1 compares the Monte Carlo estimates for selected functionals of interest based on 5000 MCMC samples from the Hamiltonian no–u–turn sampler (`R` package `rstan`), against those provided by the Monte Carlo and approximate methods discussed in Sections 4.3.2.1–4.3.2.2. In particular, we compute such functionals using both 5000 i.i.d. samples from the exact SUN posterior provided by Algorithm 6, and also by leveraging the strategies associated with the blocked partially–factorized mean–field approximation in Algorithm 7. In computing such an approximation under the sequential discrete choice multinomial probit model, we follow the guidelines in Remark 4.7 and group those augmented data corresponding to the same unit i. We shall emphasize that when the coefficients are not shared across labels and have independent priors, the overlap among rows of $\bar{\mathbf{X}}$ referring to the same unit i is absent in sequential discrete choice representations. Hence, in this very specific case, we have that $\min_{q(\boldsymbol{\beta},\bar{\mathbf{z}})\in\mathcal{Q}_{\text{PMF-B}}}\text{KL}[q(\boldsymbol{\beta},\bar{\mathbf{z}})||p(\boldsymbol{\beta},\bar{\mathbf{z}} \mid \mathbf{y},\mathbf{X})] = \min_{q(\boldsymbol{\beta},\bar{\mathbf{z}})\in\mathcal{Q}_{\text{PFM}}}\text{KL}[q(\boldsymbol{\beta},\bar{\mathbf{z}})||p(\boldsymbol{\beta},\bar{\mathbf{z}} \mid \mathbf{y},\mathbf{X})]$. As we will discuss in the following, this blocking approach is more crucial for the multinomial probit models in Sections 4.2.1–4.2.2. To highlight the benefits of the blocked partially–factorized approximation, we also compare results with classical mean–field variational Bayes assuming independence between $\boldsymbol{\beta}$ and $\bar{\mathbf{z}}$ (Consonni and Marin, 2007; Girolami and Rogers, 2006).

As highlighted in Figure 4.1, the two sampling–based methods provide comparable results in terms of inference and prediction. However, Algorithm 6 produces almost 75 samples of $\boldsymbol{\beta}$ per second, whereas the Hamiltonian no–u–turn sampler can only draw one sample every 3 seconds. This massive computational cost makes state–of–the–art MCMC methods rapidly unfeasible in large p settings. We shall highlight that by relying on i.i.d. samples, Algorithm 6 has also the advantage of avoiding the need of burn–in periods and convergence checks. However, as discussed in Section 4.3.2, Algorithm 6 scales poorly the with sample size and, hence, it becomes impractical in studies with n larger than a few hundreds. This motivates the blocked partially–factorized approximation in Section 4.3.2.2, which notably matches almost perfectly the Monte Carlo estimates in such a high–dimensional setting with $p > n$ (see Figure 4.1) and requires only 0.25 seconds to converge and 16 seconds to compute the different functionals. Classical mean–field variational

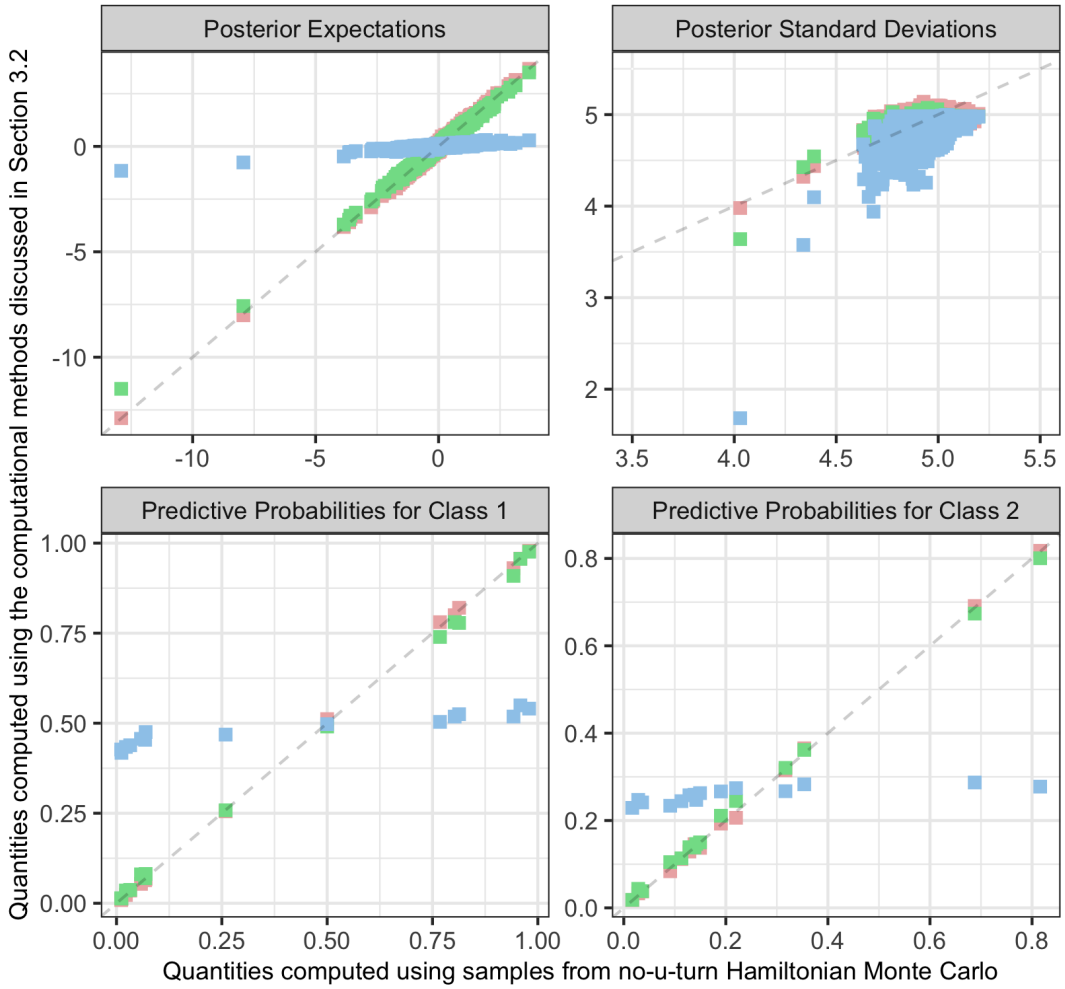

Figure 4.1: Comparison between the estimates of key functionals obtained under the methods discussed in Sections 4.3.2.1 and 4.3.2.2 (y–axis), against those provided by the `STAN` implementation of the Hamiltonian no–u–turn sampler (y–axis). Red squares refer to Monte Carlo estimates based on i.i.d. samples from the exact SUN posterior produced by Algorithm 6, whereas blue and green squares denote the estimates provided by classical mean–field variational Bayes and by our blocked partially–factorized approximation, respectively.

Bayes has comparable running times, but the independence assumption between $\boldsymbol{\beta}$ and $\bar{\mathbf{z}}$ induces notable overshrinkage of the locations and scales, which massively affects the estimation of the predictive probabilities.

Before concluding our analysis, we also implement the multinomial probit model with class–specific parameters presented in Section 4.2.2, assuming independent standard normal errors. Due to the form of the dataset, the classical discrete multinomial probit in Section 4.2.1 is not appropriate, since it would require a vector of covariates for each combination of unit i and lesion l, which is not the case for this study. Nonetheless,

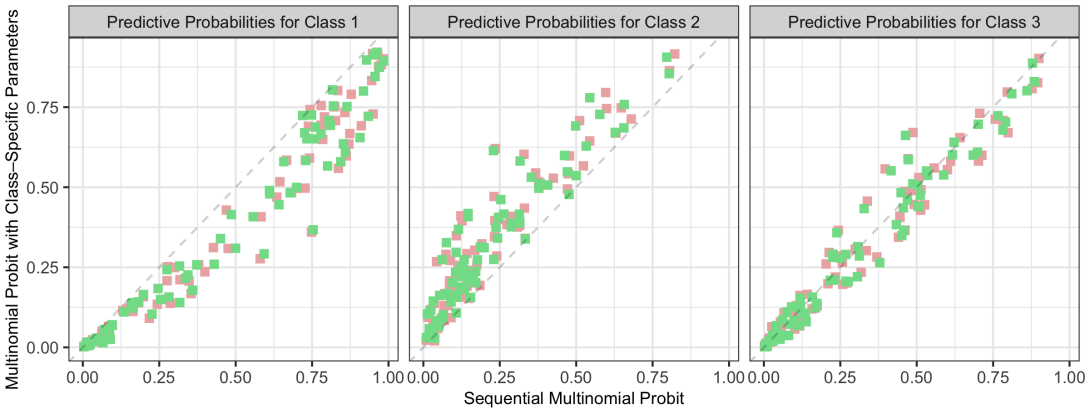

Figure 4.2: Comparison among the predictive probabilities provided by model (4.3) (y–axis) and (4.5) (x–axis). The dataset has been divided into 6 folds, and for each fold the predictive probabilities are computed using all the other available data as training set. Red squares refer to Monte Carlo estimates based on i.i.d. samples from the exact SUN posterior produced by Algorithm 6, whereas the green squares denote the estimates provided by our blocked partially–factorized approximation.

according to the results in Sections 4.2.1, 4.2.2, and 4.3, models (4.1) and (4.3) induce posteriors with comparable dimensions and, hence, the performance of the multinomial probit with class–specific coefficients is also indicative of the one associated with the classical specification in Section 4.2.1. Here, we focus on comparing the computational and predictive performance between the already–implemented sequential formulation in (4.5) and the one with class–specific coefficients in (4.3), considering the Monte Carlo and variational estimates discussed in Section 4.3.2. Under model (4.3), blocking across units i was more crucial to obtain accurate variational inference. The Hamiltonian no–u–turn sampler faced, instead, severe mixing and convergence issues under model (4.3), further highlighting major issues of MCMC in such settings.

Figure 4.2 compares variational and Monte Carlo estimates of the predictive probabilities for all the units, under the two models. To estimate the predictive probabilities we split the dataset in six folds, four having 13 observations and two having 12 observations. Then, we compute the predictive probabilities for the observations in each fold, using the units in the remaining five folds to obtain the posterior distribution. As it can be noticed from Figure 4.2, the two models provide similar, but not identical, predictive probabilities, whose values are almost the same when comparing Monte Carlo and variational estimates. This result confirms the excellent performance of the proposed blocked partially–factorized approximation in high–dimensional settings, especially when $p > n$. Indeed, by slightly increasing the dimension of the training set, the number of $\boldsymbol{\beta}$ samples per second produced by Algorithm 6 rapidly decreases from 75 to 50 in model (4.5), whereas the variational strategy still requires about 0.25 seconds to converge and 16 sec-

onds to compute the functionals. The overall out–of–sample predictive accuracy under the two models is about 66.5%. Considering the simplicity of the multinomial probit models implemented, these values are quite satisfactory when compared with the 73.68% accuracy obtained under sophisticated black–box machine learning algorithms (Mesejo et al., 2016).

4.5 Discussion

This contribution provides novel conjugacy results and computational methods for a large class of multinomial probit models (Hausman and Wise, 1978; Stern, 1992; Tutz, 1991) with Gaussian priors, and extends these properties to the entire class of SUN (Arellano-Valle and Azzalini, 2006) priors. As discussed in Sections 4.3 and 4.4, the availability of a SUN posterior allows major advances in terms of closed–form, Monte Carlo and approximate variational inference which close a still unaddressed gap of MCMC methods in large p studies. These are common settings in various fields, such as in medical applications collecting a huge number of predictors via state–of–the–art imaging technologies.

Our results open also several avenues for future research. For example, although Bayesian estimation and inference for the covariance matrix $\boldsymbol{\Sigma}$ of the errors goes beyond the scope of our contribution, such a matrix can be possibly estimated via the maximization of the closed–form marginal likelihood in Corollary 4.5. If instead $\boldsymbol{\Sigma}$ is assigned a prior and the focus is on the entire posterior distribution, it could be of interest to incorporate Algorithm 6 within the Gibbs samplers by e.g. McCulloch and Rossi (1994); McCulloch et al. (2000) and Imai and Van Dyk (2005), to improve mixing when sampling from the full–conditional $(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\Sigma})$.

The results in this chapter can be also included in more complex formulations. For instance, the sequential representation in model (4.5) has been used also within Bayesian nonparametric hierarchical models for density regression based on probit stick-breaking process (Rodriguez and Dunson, 2011). Our results could be useful in such settings to improve the computational performance and the theoretical treatment of predictor–dependent Bayesian nonparametric mixture models. Also extensions of our results to classification via Gaussian processes (Rasmussen and Williams, 2006; Girolami and Rogers, 2006) are straightforward. Finally, it would be interesting to exploit methods in Genton et al. (2018) to identify suitable blocks of augmented data in a more data–driven way, which can be applied to perform highly accurate variational inference not only in multinomial but also in binary probit regression.

4.A Appendix: Proofs

4.A.1 Proof of Theorem 4.4

To prove Theorem 4.4, it suffices to apply the Bayes rule and recognize a SUN density in the kernel of $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})$. In particular, note that $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) \propto p(\boldsymbol{\beta})p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X}) \propto \phi_q(\boldsymbol{\beta}-\boldsymbol{\xi}; \boldsymbol{\Omega})\Phi_h(\boldsymbol{\gamma}+\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}); \boldsymbol{\Gamma}-\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta})\Phi_m(\bar{\mathbf{X}}\boldsymbol{\beta}; \boldsymbol{\Lambda})$ and re–write $\Phi_m(\bar{\mathbf{X}}\boldsymbol{\beta}; \boldsymbol{\Lambda})$ as

$$\Phi_m[\mathbf{s}^{-1}\bar{\mathbf{X}}\boldsymbol{\xi} + (\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}\bar{\mathbf{X}}^\intercal\mathbf{s}^{-1})^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}); \mathbf{s}^{-1}(\bar{\mathbf{X}}\boldsymbol{\Omega}\bar{\mathbf{X}}^\intercal + \boldsymbol{\Lambda})\mathbf{s}^{-1} - \mathbf{s}^{-1}\bar{\mathbf{X}}\boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}\bar{\mathbf{X}}^\intercal\mathbf{s}^{-1}]$$

Replacing this quantity in the kernel of the posterior and leveraging known properties of Gaussian cumulative distribution functions, it follows that

$$\begin{aligned}&\Phi_h(\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta})\Phi_m(\bar{\mathbf{X}}\boldsymbol{\beta}; \boldsymbol{\Lambda})\\ &\quad = \Phi_{h+m}(\boldsymbol{\gamma}_{\text{post}}+\boldsymbol{\Delta}_{\text{post}}^\intercal\bar{\boldsymbol{\Omega}}_{\text{post}}^{-1}\boldsymbol{\omega}_{\text{post}}^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_{\text{post}}); \boldsymbol{\Gamma}_{\text{post}}-\boldsymbol{\Delta}_{\text{post}}^\intercal\bar{\boldsymbol{\Omega}}_{\text{post}}^{-1}\boldsymbol{\Delta}_{\text{post}}),\end{aligned}$$

with $\boldsymbol{\xi}_{\text{post}}$, $\boldsymbol{\Omega}_{\text{post}}$, $\boldsymbol{\Delta}_{\text{post}}$, $\boldsymbol{\gamma}_{\text{post}}$ and $\boldsymbol{\Gamma}_{\text{post}}$ as in Theorem 4.4. Leveraging this equality and recalling that $\boldsymbol{\xi}_{\text{post}} = \boldsymbol{\xi}$, $\boldsymbol{\Omega}_{\text{post}} = \boldsymbol{\Omega}$, it can be easily noticed that $p(\boldsymbol{\beta})p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X})$ coincides with the kernel of the SUN in (4.7), thus proving Theorem 4.4. To prove that $\boldsymbol{\Omega}^*_{\text{post}}$ is a correlation matrix it suffices to replace $\mathbf{I}_n$ with $\boldsymbol{\Lambda}$ in the proof of Corollary 4 in Durante (2019). □

4.A.2 Proof of Corollary 4.5

To derive equation (4.8), note that according to the proof of Theorem 4.4, $p(\boldsymbol{\beta})\Phi_m(\bar{\mathbf{X}}\boldsymbol{\beta}; \boldsymbol{\Lambda}) = p(\mathbf{y}, \boldsymbol{\beta} \mid \mathbf{X}) = p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})\Phi_{h+m}(\boldsymbol{\gamma}_{\text{post}}; \boldsymbol{\Gamma}_{\text{post}})/\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})$. Hence,

$$\begin{aligned}p(\mathbf{y} \mid \mathbf{X}) &= [p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})\Phi_{h+m}(\boldsymbol{\gamma}_{\text{post}}; \boldsymbol{\Gamma}_{\text{post}})/\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})]/p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})\\ &= \Phi_{h+m}(\boldsymbol{\gamma}_{\text{post}}; \boldsymbol{\Gamma}_{\text{post}})/\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma}).\end{aligned}$$

□

4.A.3 Proof of Corollary 4.6

To prove Corollary 4.6 simply notice that equation (4.9) is the ratio between the marginal likelihoods of the expanded dataset and the observed one—without the additional unit with response $y_{\text{new}} = l$ and covariates $\mathbf{x}_{\text{new}}$. Hence, the expression for the predictive probabilities follows directly from Corollary 4.5 after noticing that, due to the conditional independence assumption in (4.1), (4.3) or (4.5), one has $p(\mathbf{y} \mid \mathbf{X}, \mathbf{x}_{\text{new}}) = p(\mathbf{y} \mid \mathbf{X})$. □

Chapter 5

The Hidden Hierarchical Pitman-Yor Process

5.1 Introduction

Species sampling models are a popular tool to face one of the most important problems in Statistics: prediction. They owe their name to the seminal contributions by Good (1953) and Good and Toulmin (1956), which originated from ecological applications, and were first studied from a Bayesian nonparametrics perspective in Lijoi et al. (2007). In such a setting, the main interest is in prediction of additional observations, conditionally on the available data, with particular focus on the number of new species in an additional sample, which can be seen as a measure of species diversity, or the rate of decay of the probability of discovering new species. Since the original formulation, the term 'species sampling model' has been broadly used for a wide range of discrete distributions, not necessarily linked to the initial ecological and biological applications, while maintaining the original terminology and denoting as 'species' the unique values that the observations can take (Pitman, 1996). The term 'species' has then gained a metaphoric meaning which can change depending on the context, denoting, for instance, different possible types, genes, agents or categories,

Lately, species sampling models faced a growing interest from both applied and theoretical perspective, with applications in several fields such as genetics (Lijoi et al., 2007; Favaro et al., 2009, 2012), economics (Lijoi et al., 2016), and machine learning (Teh, 2006) just to mention a few. See also De Blasi et al. (2015) for an extensive overview on their use in the Bayesian nonparametrics framework and other possible applications. In this Bayesian setting, these constructions have been further generalized to effectively tackle the problem of prediction when the data arise from different related experiments or populations, i.e. when we are in the so-called partially exchangeable framework. In such a scenario, Bayesian hierarchical models can be successfully applied to naturally borrow

information across the different populations to improve the predictive performance of the model. This is the underlying idea of some of the most popular Bayesian nonparametrics constructions as the hierarchical and nested formulations for the Dirichlet Process (DP) (Ferguson, 1973) and their generalizations to the Pitman-Yor process (PYP) and beyond (Teh, 2006; Teh et al., 2006; Rodríguez et al., 2008; Camerlenghi et al., 2017, 2019b).

Despite the availability of a large number of contributions in the literature to face the species sampling problem in a single population framework, just a few of them consideres the more challenging case of multiple populations. Battiston et al. (2018) and Camerlenghi et al. (2017) exploit a hierarchical Pitman-Yor process (HPYP) construction to effectively face the problem of prediction combining different populations. The choice of the HPYP arises naturally in the species-sampling framework, as the random partition structure induced by the PYP is governed by two parameters and is such that the probability of observing a new species for an additional observation depends on the number of distinct species observed so far, while in the DP case there is only one parameter governing the clustering structure and the above mentioned probability depends only on the global sample size.

This different behaviour gives rise to different asymptotic distributions for the number of observed clusters as the population size diverges, with the PYP showing a power-law behavior, which is observed in many empirical studies (Mitzenmacher, 2004; Goldwater et al., 2006), while the DP shows only a logarithmic growth, which appears too restrictive. However, the hierarchical construction exploited in the two above-mentioned contributions does not allow to naturally test homogeneity of subpopulations and cluster the populations with the same *species* distributions. Motivated by the above-mentioned issues we define a novel hierarchical construction based on PYPs which allows to effectively face also the aforementioned task. This model is obtained by adding a latent nonparametric discrete prior distribution on the population distributions, so that ties among them are allowed. In such a setting, testing for homogeneity of population distributions arises naturally, as the model allows to perform probabilistic clustering of the distributions of the groups.

5.2 Preliminaries

Before presenting the proposed model in Section 5.3, we shortly review the literature involved in such construction. Following Pitman (1996), a random probability P is said to be distributed according to a proper species sampling process if it admits the series representation

$$P = \sum_{i \geq 1} \pi_i \delta_{X_i^*}, \quad (X_i^*)_{i \geq 1} \overset{iid}{\sim} H \perp (\pi_i)_{i \geq 1},$$

with H non-atomic. The law of P is completely specified after one fixes the law of the vector of weights $(\pi_i)_{i \geq 1}$. In particular, when the π_i's are such that $\pi_i = v_i \prod_{l=1}^{i-1} v_l$, with

$v_i \sim \text{BETA}(1-\sigma, \theta+i\sigma)$, $i \geq 1$, $\sigma \in [0,1)$ and $\theta > -\sigma$, then P is distributed according to a PYP with parameters (θ, σ, H), denoted $P \sim \text{PYP}(\theta, \sigma; H)$. This process is also called two-parameter Poisson-Dirichlet process, and its particular case $\sigma = 0$ boils down to the DP. Observe that, although in species sampling processes the base measure H is nonatomic, in the general PYP formulation this is not required. A vector of weights $(\pi_i)_{i\geq 1}$ constructed with the process just described is said to be $\text{GEM}(\sigma, \theta)$ distributed, after Griffiths, Engen, and McCloskey. A well-known urn scheme allows to sequentially sample observations from P since if $\boldsymbol{U}_n = (U_1, \ldots, U_n)$ is a conditionally independent sample from P, i.e. $U_i \mid P \overset{iid}{\sim} P$, then a new observation U_{n+1} will have predictive distribution

$$U_{n+1} \mid \boldsymbol{U}_n \sim \sum_{i=1}^{K_n} \frac{n_i - \sigma}{\theta + n} \delta_{U_i^*}(\cdot) + \frac{\theta + K_n \sigma}{\theta + n} H(\cdot), \tag{5.1}$$

where K_n is the number of distinct values $(U_1^*, \ldots, U_{K_n}^*)$ in the sample $\boldsymbol{U}_n$, and n_i are their multiplicities, so that $\sum_{i=1}^{K_n} n_i = n$.

This single–sample scenario is well established in the literature (see De Blasi et al. (2015) for a review), however in many applications the data are collected in J different, but related, experiments or populations. In the following we denote with $\boldsymbol{X} = \{(X_{j,i})_{i\geq 1} : j = 1, \ldots, J\}$ the data matrix. In such a framework the assumption of a common underlying distribution (*exchangeability*) is too restrictive since it does not take into account the possible differences of the populations. On the other hand, the assumption of independence across populations does not allow to borrow information across experiments in the Bayesian learning.

A natural compromise between the aforementioned extreme cases is partial exchangeability (de Finetti, Bruno, 1938), that entails exchangeability within but not across the different groups. Thanks to de Finetti's theorem, we can characterize the array $\boldsymbol{X}$ as arising from a vector of J dependent random probabilities. More precisely, for every vector of population sample sizes $(I_1, \ldots, I_J)$, it holds

$$\begin{aligned} X_{j,i} \mid (P_1, \ldots, P_J) &\sim P_j \quad (i = 1, \ldots, I_j; j = 1, \ldots, J) \\ (P_1, \ldots, P_J) &\sim \mathcal{L}, \end{aligned}$$

where $\mathcal{L}$ takes the role of the prior in the Bayes-Laplace paradigm and controls the dependence, thus the borrowing of information, across the different populations.

Many possible prior specifications for the vector $(P_1, \ldots, P_J)$ are possible. When dealing with species sampling problems, one of the most famous priors in a single–population framework is arguably the PYP. This is due to the fact that, as apparent from equation (5.1), when sampling a new out-of-sample observation, the probability to allocate it to a new cluster depends on the number of already created cluster, and not only on the total number of observations, as happens instead in the case of a DP prior. For this reason, together with the asymptotic power law shown by the number of clusters as n diverges,

the PYP is usually the first choice in species sampling problems, being the DP a valuable choice for density estimation under mixture models, but not flexible enough for species sampling processes. Consistently, a common prior specification in multiple-sample cases for $(P_1, \dots, P_J)$ is the HPYP (Teh, 2006; Teh et al., 2006; Battiston et al., 2018; Camerlenghi et al., 2017). This construction is shortly reviewed in Section 5.2.1: although being well-suited for multiple-sample prediction, it does not allow to test for distribution homogeneity across different populations. This is one of the two tasks of interest in the present chapter, and, to the best of the authors' knowledge, its treatment in the species sampling framework is lacking, aside from early attempts by Lijoi et al. (2008). In order to achieve this, a nested structure is added, allowing for possible ties in the group distributions P_j. This is done exploiting a nested Pitman-Yor process (NPYP), which is introduced in Section 5.2.2 and follows from the nested Dirichlet Process (NDP) (Rodríguez et al., 2008), after replacing the DP with a PYP.

5.2.1 Hierarchical Pitman-Yor process

A well-known Bayesian nonparametric prior for a vector of dependent discrete random probabilities $(P_1, \dots, P_J)$ is given by the hierarchical Pitman-Yor process (HPYP) (Teh, 2006; Teh and Jordan, 2010), which extends the definition of the hierarchical DP (Teh et al., 2006).

The idea is to introduce dependence across the random probabilities $P_1, \dots, P_J$ via a common random discrete base measure P_0. More precisely we say that $(P_1, \dots, P_J)$ follows a HPYP with parameter vector $(\sigma, \theta, \sigma_0, \theta_0, H)$, denoted $(P_1, \dots, P_J) \sim \text{HPYP}(\sigma, \theta, \sigma_0, \theta_0; H)$ if

$$P_j \mid P_0 \overset{iid}{\sim} \text{PYP}(\sigma, \theta; P_0) \;\; j = 1, \dots, J, \qquad P_0 \sim \text{PYP}(\sigma_0, \theta_0; H).$$

Thanks to the discreteness of P_j we will observe ties with positive probability between the observations recorded in each population $\boldsymbol{X}_j = \{X_{j,i} : i = 1, \dots, I_j\}$. Furthermore, the discreteness of the common random base measure P_0 allows to share species (cluster observations) across the random probabilities. This feature is essential to perform clustering with mixture models as well as species sampling under heterogeneous populations (Teh et al., 2006; Camerlenghi et al., 2017).

This random partition structure induced by the ties is the core element of species sampling models and from a statistical perspective it can be interpreted as a random clustering. The probability distribution of such a random partition structure can be characterized via the *partial exchangeable partition probability function* (pEPPF) marginalizing out the vector of random probabilities. The pEPPF is an essential object to understand the model and perform inference. For instance, from the pEPPF we can derive closed-form results for the joint moments of the observations, both in the same or different populations.

Moreover, it can also be used to derive urn schemes that allow to develop marginal MCMC routines which constitute the basis to perform predictive inference. See Camerlenghi et al. (2019b) for results on the pEPPF for a large class of models.

However, when the goal is to test population homogeneity, the HPYP has a huge drawback, as it does not allow two groups to share the same distribution. Indeed, in the HPYP, $\text{pr}(P_j \neq P_k) = 0$ for any $j \neq k$. In order to allow for homogeneous subgroups of populations we will rely on nested structures, extending the HPYP in order to allow $P_j = P_k$, for $j \neq k$, with positive probability. Thus, before moving to the presentation of the proposed model, we introduce the nested Pitman-Yor process (NPYP).

5.2.2 Nested Pitman-Yor process

The nested Dirichlet process (NDP) (Rodríguez et al., 2008) is arguably the most famous Bayesian nonparametric prior to perform joint clustering of distributions and observations under mixture models. However, as pointed out by Camerlenghi et al. (2019a) it suffers from a *degeneracy issue* that makes it unsuitable to face our species sampling problem. More precisely, it allows to naturally test for homogeneity of groups and to perform probabilistic clustering of groups since, contrary to the HDP case, a priori we have $\text{pr}(P_j \neq P_k) \in (0, 1)$, for any $j \neq k$. However, given that a single *species* (cluster of observations) is shared across groups j and k, i.e. $X_{j,i} = X_{k,l}$ for some $i, l \geq 1$, the species-populations P_j and P_k are almost surely equal. On the other hand, given that the two species-populations are not exactly equal they are independent and cannot share any species.

In order to overcome the restrictions not suitable for species sampling problems due to a DP prior exposed in Section 5.2, we first extend the hierarchical definition of the NDP to a composition of PYPs. However, also such nested Pitman-Yor process (NPYP) suffers from the same *degeneracy issue* of the NDP. This will be overcome in Section 5.3, where we introduce a novel prior for dependent species sampling processes that solves the issue combining the NPY and the HDP, taking the advantages of the both of them.

We say that $(P_1, \ldots, P_J)$ follows a NPYP distribution with vector of parameters $(\alpha, \gamma, \sigma, \theta, H)$, denoted $(P_1, \ldots, P_J) \sim \text{NPYP}(\alpha, \gamma, \sigma, \theta; H)$, if

$$P_j \mid Q \overset{iid}{\sim} Q \quad j = 1, \ldots, J, \qquad Q \sim \text{PYP}(\alpha, \gamma; \text{PYP}(\sigma, \theta; H)). \tag{5.2}$$

In order to ease the understanding of the model we can rewrite the random distribution on the space of distributions Q exploiting the well-known stick-breaking representation of the Pitman-Yor process, so that

$$Q = \sum_{k \geq 1} \omega_k^* \delta_{P_k^*},$$

where the unique atoms P_k^* are random probabilities on the space of the observations and are i.i.d. samples from $\text{PYP}(\sigma, \theta; H)$, independent of the weights $(\omega_k^*)_{k \geq 1} \sim \text{GEM}(\alpha, \gamma)$.

The discreteness of Q induces a probabilistic clustering of the groups since $\text{pr}(P_j = P_k) = \frac{1-\alpha}{\gamma+1} \in (0,1)$. However, as for the NDP, given that a single atom is shared between the two distributions, such probability to degenerate to the exchangeable case is 1. Indeed, given that the two distributions P_j and P_k are different they are i.i.d. sampled from $\text{PYP}(\sigma, \theta; H)$ and thus their random atoms are i.i.d. sampled from a non-atomic distribution H and are almost surely different.

To overcome such an issue of the NDP in mixture models Camerlenghi et al. (2019a) introduce a novel class of Bayesian nonparametrics priors named latent nested processes (LNPs). LNPs have the merit to be the first proposal to solve the degeneracy issue of the NDP. However, they are not suited for the study at hand, since computations become infeasible when there are more than two groups and in addition they force the proportion of species, i.e. the weights, to be the same across groups.

Other proposals are available in the literature, exploiting hidden hierarchical Dirichlet process (HHDP) constructions for mixture models. However, in addition to having a different focus, the theoretical results in such frameworks as well the proposed algorithms are not suited for the scenario we are considering, since they rely on the conjugacy and the finite dimensional approximations of the DP. See also Soriano and Ma (2019), Christensen and Ma (2020) and Beraha et al. (2020) for stimulating contributions to this literature. Notice that, even if for practical reason we restrict ourselves to the case of composition of PYPs, the methodological arguments together with the algorithms developed in the present work can be easily adapted to a more general class of priors that arise from the composition of different Gibbs type priors, due to product form of their exchangeable partition probability function (EPPF).

5.3 Hidden hierarchical Pitman-Yor process

After having addressed the limitations of the HPYP and NPYP for the scopes at hand, we introduce a novel class of priors, called hidden hierarchical Pitman–Yor process (HHPYP), arising from a composition of PYPs that overcomes the above mentioned issues. In particular, this construction is obtained combining the HPYP with the NPYP, as explained in Section 5.3.1, and allows for ties in the population distributions, without suffering from the aforementioned *degeneracy* issue of the NPYP, thus making homogeneity testing of sub-groups effective, while simultaneously performing species sampling tasks borrowing information across populations.

5.3.1 Definition and basic properties

The HHPYP is obtained by taking a NPYP with discrete base measure distributed according to a PYP. This hierarchical construction allows different populations P_j and P_k,

$j \neq k$, to possibly share the same atoms, so that a tie in two observations in these groups does not imply $P_j = P_k$ with probability 1.

In formulae, we say that $(P_1, \ldots, P_J) \sim \text{HHPYP}(\alpha, \gamma, \sigma, \theta, \sigma_0, \theta_0; H)$ if

$$\begin{aligned} (P_1, \ldots, P_J) &\sim \text{NPYP}(\alpha, \gamma, \sigma, \theta; P_0^*) \\ P_0^* &\sim \text{PYP}(\sigma_0, \theta_0; H). \end{aligned} \tag{5.3}$$

From now on we assume that the common probability on the sample space H is non-atomic and for notational simplicity we just write $(P_1, \ldots, P_J) \sim \text{HHPYP}$. Furthermore, we assume the hyperparameters to be fixed, but in practice we can set a prior on them and all the results hold given the hyperparameters and it is straightforward to adapt the Gibbs sampler in Section 5.4 as for the usual species sampling under PYP prior in the exchangeable case.

It follows from (5.2) that we can alternatively characterize the P_j's to be i.i.d. sampled from $Q \sim \text{PYP}(\alpha, \gamma; \text{PYP}(\sigma, \theta; P_0^*))$, given P_0^*, which admits the representation

$$Q = \sum_{k \geq 1} \omega_k^* \delta_{P_k^*},$$

where the weights $(\omega_k^*)_k \sim \text{GEM}(\alpha; \gamma)$ are independent from the distribution atoms. The unique underlying distributions $(P_k^*)_{k \geq 1}$ follow an infinite dimensional HPYP, that is

$$P_k^* \mid P_0^* \overset{iid}{\sim} \text{PYP}(\theta, \sigma; P_0^*) \ \ (k \geq 1), \quad P_0^* \sim \text{PYP}(\theta_0, \sigma_0; H).$$

The discreteness of Q allows to cluster the distributions. For instance, $\text{pr}(P_j = P_k) = \frac{1-\alpha}{\gamma+1} \in (0, 1)$, as for the NPYP. However, thanks to the discreteness of the common random base measure P_0^* the unique random distributions P_k^*'s are now dependent and share the same countable set of atoms allowing to share species across populations which is essential to overcome the aforementioned *degeneracy* issue.

In order to better understand the model, the role of the hyperparameters and the borrowing of strength we can derive the moments of the random probability measures $(P_1, \ldots, P_J) \sim \text{HHPYP}$ evaluated at an arbitrary measurable set A of the sample space $\mathbb{X}$. All the proofs are available in the appendix. The expected value is $\mathbb{E}[P_j(A)] = H(A)$, as usual in species sampling processes, while the variance can be derived leveraging results on hierarchical models (Camerlenghi et al., 2019b) and has the form

$$\text{Var}[P_j(A)] = \frac{H(A)[1 - H(A)]}{\theta_0 + 1} \left[(1 - \sigma_0) + (\theta_0 + \sigma_0) \frac{1 - \sigma}{\theta + 1} \right]. \tag{5.4}$$

We can also derive the expression for the correlation between P_j and P_k, $j \neq k$, which does not depend on the specific set A, and thus is often interpreted as a global measure of dependence between the random probabilities in Bayesian nonparametrics. It holds

$$\text{Cor}[P_j(A), P_{j'}(A)] = \frac{1 - \alpha}{\gamma + 1} + \frac{\gamma + \alpha}{\gamma + 1} \frac{1 - \sigma_0}{(1 - \sigma_0) + (\theta_0 + \sigma_0) \frac{1 - \sigma}{\theta + 1}}. \tag{5.5}$$

It is interesting to notice the role played by the parameters α and γ, with the correlation decreasing as $\alpha \to 1$ or $\gamma \to \infty$: this is indeed consistent with the fact that in such scenarios we are decreasing the probabilities of homogeneity between the two populations. However, contrary to the NPYP (and its special case NDP), if $j \neq k$, P_j and P_k are not independent, but follow a bi-dimensional HPYP and we can control their dependence via the hyperparameters $(\sigma, \theta, \sigma_0, \theta_0)$ as for the well-known HPYP.

Finally, if the focus is predict future observations it is better to study the dependence directly in term of the observable random variables as de Finetti suggested. If the data matrix $\boldsymbol{X}$ is drawn from $(P_1, \ldots, P_J) \sim$ HHPYP, then

$$\begin{aligned}\mathrm{Cor}(X_{j,i}, X_{k,l}) &= \mathrm{pr}(X_{j,i} = X_{k,l}) \\ &= \begin{cases} \left[\left(\frac{1-\sigma}{\theta+1} + \frac{1-\sigma_0}{\theta_0+1}\frac{\theta+\sigma}{\theta+1}\right)(1-\alpha) + \frac{1-\sigma_0}{\theta_0+1}(\gamma+\alpha)\right](\gamma+1)^{-1} & \text{if } j \neq k \\ \left[1-\sigma+\frac{1-\sigma_0}{\theta_0+1}(\theta+\sigma)\right](\theta+1)^{-1} & \text{if } j = k. \end{cases} \end{aligned} \tag{5.6}$$

Notice that a priori the correlation between observations, i.e. the probability that the observations belong to the same species, is larger when they arise from the same population, which is an appealing feature from a modeling perspective. The fact that the correlation between two observations coincides with the probability that they are equal is a very general result for species sampling models, both in the exchangeable and partially exchangeable cases. See the proof in the Appendix for further insights.

This hierarchical representations of general dependent species sampling processes points out that the dependence is controlled by the ties of the observations and the random partitions they induce. Thus, in order to understand the model and develop sampling schemes, we now study the random partitions structures of the distributions and populations induced by the ties.

5.3.2 Partially Exchangeable Partition Probability Functions and Urn Schemes

A priori, the discreteness of Q induces a random partition $\Psi^{(J)}$ of $[J] = \{1, \ldots, J\}$ and thus a clustering of the distributions $P_1, \ldots, P_J$. More precisely, if $(P_1, \ldots, P_J) \sim$ HHPYP the probability law of $\Psi^{(J)}$ is characterized by the following EPPF, arising from the PYP,

$$\phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma) = \frac{\prod_{r=1}^{R-1}(\gamma + r\,\alpha)}{(\gamma+1)_{J-1}} \prod_{r=1}^{R}(1-\alpha)_{m_r-1}, \tag{5.7}$$

where $(x)_J = x(x+1)\cdots(x+J-1)$ is the Jth ascending factorial, R is the random number of blocks of the partition of $[J]$ and m_r is the cardinality of the rth block in order of arrival of the unique P_j's. Equation (5.7) immediately follows after recognizing that

the underlying distributions $P_1^*, \dots, P_R^*$ are almost surely different under the HHPYP, although they can share the same atoms.

Denoting with $\boldsymbol{S} = (S_1, \dots, S_J)$, $S_1 = 1$, the cluster membership indicator vector of the J populations in the Chinese restaurant process (CRP), the following Pölya urn scheme characterizes the distribution of $\boldsymbol{S} = (S_1, \dots, S_J)$:

$$\text{pr}(S_{j+1} = S \mid S_1, \dots, S_j) = \begin{cases} \frac{m_r^{-(j+)} - \gamma}{m_\cdot + J} & \text{if } S = r, \text{ for } r = 1, \dots, R^{-(j+)}, \\ \frac{\alpha + \gamma R^{-(j+)}}{m_\cdot + J} & \text{if } S = R^{-(j+)} + 1, \end{cases} \tag{5.8}$$

where we use the $\cdot$ symbol to indicate a summation over an index set, $(j+) = (j+1, \dots, J)$ is the set of future populations not assigned to any restaurant yet, and $a^{-(b)}$ denotes the quantity a without considering the elements in b.

In addition, the discreteness of the P_j's induces a random partition of the observations $\boldsymbol{X}$ within and across populations. Calling D the overall number of unique values (number of species) in $\boldsymbol{X}$ and $\boldsymbol{n}_j = (n_{j,d} : d = 1, \dots, D)$ the vector of cardinalities of the species observed in population j, $j = 1, \dots, J$, the above mentioned partition structure of $\boldsymbol{X}$ is characterized by the pEPPF $\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J)$. In order to have a tractable form for it, in addition to the population assigment vector $\boldsymbol{S}$, we also make use of a further data augmentation, which corresponds to the usual table augmentation of the Chinese restaurant franchise (CRF) (Teh, 2006; Teh and Jordan, 2010). More precisely, exploiting that culinary metaphor, we introduce the variables $T_{j,i}$, $j = 1, \dots, J$, $i = 1, \dots, I_j$, representing the table at which observation i in population j sits and denote $\boldsymbol{T} = \{T_{j,i} : \ j = 1, \dots, J, \ i = 1, \dots, I_j\}$. Furthermore, we call $q_{r,t,d}$ the number of customers in restaurant r sitting at table t eating dish d. Marginalizing out the previous latent variables we obtain the following form for the pEPPF.

Theorem 5.3.1. *If $\boldsymbol{X}$ is drawn from $(P_1, \dots, P_J) \sim \text{HHPYP}(\alpha, \gamma, \sigma, \theta, \sigma_0, \theta_0; H)$, then the random partition structure induced by the samples is characterized by the following pEPPF*

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J) = \sum \phi_R^{(J)}(m_1, \dots, m_R; \alpha, \gamma) \Phi_D^{(n)}(q_{1,\cdot,\cdot}, \dots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0),$$

where the sum is over all partitions of $[J]$, $\phi_R^{(J)}$ as in (5.7), and $\Phi_D^{(n)}(q_{1,\cdot,\cdot}, \dots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0)$ is the pEPPF associated to an R-dimensional $\text{HPYP}(\sigma, \theta, \sigma_0, \theta_0; H)$.

Exploiting the aforementioned variable augmentation based on $\boldsymbol{T}$ and $\boldsymbol{S}$, and calling $X_1^*, \dots, X_D^*$ the unique values in the sample $\boldsymbol{X}$, it follows from Bayes Theorem that the following urn scheme easily allows to sample from (5.3) in two steps:

(1) Assign the population to the different restaurant recursive from equation (5.8).

(2) Given the assignment of the populations to the restaurants via $\boldsymbol{S}$, sample the table assignments $\boldsymbol{T}$ and the observations values $\boldsymbol{X}$ recursively adapting the CRF (Teh,

2006) from

$$\text{pr}(X_{j,i} = x, T_{j,i} = t \mid \boldsymbol{S}, \boldsymbol{X}^{-(ji+)}, \boldsymbol{T}^{-(ji+)}) =$$
$$\begin{cases} \dfrac{\theta_0 + D^{-(ji+)}\sigma_0}{\theta_0 + \ell_{\cdot,\cdot}^{-(ji+)}} \dfrac{\theta + \ell_{r,\cdot}^{-(ji+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(ji+)}} \text{ if } x = \text{“new”} \text{ and } t = \text{“new”}, \\ \dfrac{\omega_d^{-(ji+)}}{\theta_0 + \ell_{\cdot,\cdot}^{-(ji+)}} \dfrac{\theta + \ell_{r,\cdot}^{-(ji+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(ji+)}} \text{ if } x = X_d^{*-(ji+)} \text{ and } t = \text{“new”} \text{ for } d = 1, \ldots, D^{-(ji+)}, \\ \dfrac{q_{r,t,d}^{-(ji+)} - \sigma}{\theta + q_{r,\cdot,\cdot}^{-(ji+)}} \text{ if } x = X_d^{*-(ji+)} \text{ and } t = T_{r,d,l}^{*} \text{ for } l = 1, \ldots, \ell_{r,d}^{-(ji+)},\ d = 1, \ldots, D^{-(ji+)}, \end{cases}$$

where $(ji+) = \{(jl) : l \geq i\} \cup \{(kl) : k \geq j\}$ is the index set associated to the future random variables not sampled yet, and $T_{r,d,l}^{*}$ denotes the value of the lth table in restaurant r serving dish d. Finally, $\ell_{r,d}$ represents the number of tables in restaurant r serving dish d. If we are interested not just in the clustering structure, but also in the specific values of the observations, we can sample the “new” values of the observations from the non-atomic base distribution H.

Notice that, contrary to the usual CRF characterizing the HPYP, a restaurant is not identified by a unique population, but different populations can be assigned to the same restaurant, thus sharing tables. On the other hand, if two populations are assigned to two different restaurants, they will not share any table. Since this urn scheme naturally extends the well-known CRF metaphor, with the additional property that a restaurant can be composed by more than one group, we call such a Pölya urn scheme hidden Chinese restaurant franchise (HCRF). Populations are clustered together when assigned to the same restaurant. In testing the homogeneity among different groups, one can then compute the posterior probability that two populations belong to the same cluster as discussed in the next section.

5.3.3 Population homogeneity testing

One of the main goals of the present chapter is to introduce a valuable model that, among usual inferential species sampling tasks, is able to assess which populations are homogeneous. Since the clustering is probabilistic, the key quantity of interest is the posterior probability of co-clustering for each couple of distributions P_j, P_k, $j \neq k$, namely $\text{pr}(P_j = P_k \mid \boldsymbol{X})$. These probabilities can be interpreted in terms of posterior evidence of homogeneity between the distributions P_j and P_k. Considering the case of $J = 2$ populations for ease of interpretation, and denoting $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ the vectors of the counts of the overall distinct D values in each of the two populations, the pEPPF characterising the law of $\boldsymbol{X}$ can be written as

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \frac{1-\alpha}{\gamma+1}\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0) + \frac{\alpha+\gamma}{\gamma+1}\Phi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0), \quad (5.9)$$

with $\Phi_D^{(n)}$ as in Theorem 5.3.1. As expected by the model specification (5.3), the pEPPF (5.9) can be seen as a convex combination of the probability laws of the random partitions induced by different HPYPs, the first composed by a single population with $\boldsymbol{n}_1+\boldsymbol{n}_2$ vector of multiplicities, while the second formed by two distinct populations having multiplicity vectors $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ respectively. From (5.9) one can easily derive the posterior probability to degenerate to the exchangeable case, that is of the event $\{P_1 = P_2\}$.

Proposition 5.3.2. *If $J = 2$ and $\boldsymbol{X}$ is sampled from $(P_1, P_2) \sim$ HHPYP, then the posterior probability of degeneracy is*

$$\text{pr}(P_1 = P_2 \mid \boldsymbol{X}) = \frac{(1-\alpha)\Phi_D^{(n)}(\boldsymbol{n}_1+\boldsymbol{n}_2;\sigma,\theta,\sigma_0,\theta_0)}{(1-\alpha)\Phi_D^{(n)}(\boldsymbol{n}_1+\boldsymbol{n}_2;\sigma,\theta,\sigma_0,\theta_0)+(\alpha+\gamma)\,\Phi_D^{(n)}(\boldsymbol{n}_1,\boldsymbol{n}_2;\sigma,\theta,\sigma_0,\theta_0)}.$$

Notice that the HHPYP overcomes the degeneracy issue of the NDP and the NPYP allowing for the presence of shared species across populations, without implying to degenerate to exchangeability.

The above-mentioned task is strictly related with hypothesis testing procedures. Indeed, assessing whether $P_1 = P_2$, can be rephrased as a test where

$$H_0 :\ S_1 = S_2 \quad vs. \quad H_1 :\ S_1 \neq S_2. \tag{5.10}$$

H_0 and H_1 specify two different models for the data matrix $\boldsymbol{X}$. The correspondig Bayes factor is then readily available and has the form

$$B_{01} = \frac{p(\boldsymbol{X} \mid H_0)}{p(\boldsymbol{X} \mid H_1)} = \frac{\Phi_D^{(n)}(\boldsymbol{n}_1+\boldsymbol{n}_2;\sigma,\theta,\sigma_0,\theta_0)}{\Phi_D^{(n)}(\boldsymbol{n}_1,\boldsymbol{n}_2;\sigma,\theta,\sigma_0,\theta_0)}.$$

For $J > 2$, the co-clustering posterior probabilities for each couple (j, k) can be easily computed via the marginal Gibbs sampler described in Section 5.4. It will be sufficient to count how many times out of the B Gibbs updates $S_j = S_k$ to get an MCMC estimate of $\text{pr}(P_j = P_k \mid \boldsymbol{X})$. Moreover, the testing procedure (5.10) can be straightforwardly extended to the generic null hypothesis

$$H_0 :\ S_{j_1} = S_{k_1}, \ldots, S_{j_C} = S_{k_C},\ \text{for some } \{j_1, \ldots j_C\}, \{k_1, \ldots k_C\} \subseteq [J],$$

with complementary alternative hypothesis H_1. In such a case, the corresponding Bayes factor follows by specifying the summation in (5.9) to the cases specified by the null hypothesis and the alternative.

5.3.4 Inference on the number of species

Consistently with the above, let D be the overall random number of species (dishes) in the sample $\boldsymbol{X}$ of size $n = \sum_{j=1}^{J} I_j$, and call R the number of heterogeneous populations

among the J populations. To keep the notation lighter, we suppress the dependence on n, J and $q_{r,\cdot,\cdot}$. The probabilistic behavior of D and R both on finite samples and when the overall numbers of observations n and populations J diverge is of utmost importance to deeper understand key properties of the proposed species sampling model.

First, notice that $(T_{j,i} \mid S_j = r, P_r^*) \overset{iid}{\sim} P_r^*$, with $P_r^* \overset{iid}{\sim} \text{PYP}(\sigma, \theta, H)$, where H is a non-atomic probability measure, so that, if we call L_r the number of distinct values in $\boldsymbol{T}_r = (T_{j,i} : S_j = r)$, $r = 1, \ldots, R$, these quantities are independent across restaurants.

We also denote by $D_{0,\ell}$ the random number of distinct values between ℓ exchangeable values generated from P_0^*. Notice that the distribution of R, L_r and $D_{0,\ell}$ can be derived via marginalization from the EPPF induced by a PYP with non-atomic base measure. More precisely,

$$\begin{aligned} p(R) &= \frac{1}{R!} \sum_{\boldsymbol{m} \in \mathcal{T}_R(J)} \binom{J}{m_1, \ldots, m_R} \phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma) \\ &= \frac{\prod_{r=1}^{R-1}(\gamma + r\,\alpha)}{(\gamma+1)_{J-1}} \frac{\mathscr{C}(J, R; \alpha)}{\alpha^R}, \end{aligned} \tag{5.11}$$

where $\mathcal{T}_R(J) = \{(m_1, \ldots, m_R) : m_r \geq 1, \sum_{r=1}^R m_R = J\}$. Here $\mathscr{C}(n, k; \sigma)$ represents the generalized factorial coefficient $(\sigma t)_n = \sum_{k=1}^n \mathscr{C}(n, k; \sigma)(t)_k$ and computable as $\mathscr{C}(n, k; \sigma) = \frac{1}{k!} \sum_{j=0}^k (-1)^j \binom{k}{j} (-\sigma j)_n$ with the proviso $\mathscr{C}(0, 0; \sigma) = 1$ and $\mathscr{C}(n, 0; \sigma) = 1$ for any $n > 0$ and $\mathscr{C}(n, k; \sigma) = 0$ for any $k > n$. For an exhaustive review of the generalized factorial coefficients see Charalambides (2002).

Marginalizing out the corresponding EPPF we can also obtain:

$$\begin{aligned} p(D_{0,\ell}) &= \frac{\prod_{d=1}^{D_{0,\ell}-1}(\theta_0 + d\,\sigma_0)}{(\theta_0+1)_{\ell-1}} \frac{\mathscr{C}(\ell, D_{0,\ell}; \sigma_0)}{\sigma_0^{D_{0,\ell}}}, \\ p(L_r) &= \frac{\prod_{\ell=1}^{L_r-1}(\theta + \ell\,\sigma)}{(\theta_0+1)_{\ell-1}} \frac{\mathscr{C}(q_{r,\cdot,\cdot}, L_r); \sigma)}{\sigma^{L_r}}. \end{aligned}$$

In the next Theorem we derive probability distribution of the overall number of species.

Theorem 5.3.3. *If the data matrix $\boldsymbol{X}$ is drawn from $(P_1, \ldots, P_J) \sim$ HHPYP, then*

$$\begin{aligned} p(D) &= \sum_{\boldsymbol{B} \in \rho(J)} \phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma) \sum_{L=D}^n pr(D_{0,L} = D)\, pr\left(\sum_{j=1}^J L_r = L\right) \\ &= \sum_{\boldsymbol{B} \in \rho(J)} \frac{\prod_{r=1}^{R-1}(\gamma + r\,\alpha)}{(\gamma+1)_{J-1}} \prod_{r=1}^R (1-\alpha)_{m_r-1} \\ &\times \sum_{L=D}^n \frac{\prod_{d=1}^{D-1}(\theta_0 + d\,\sigma_0)}{(\theta_0+1)_{\ell-1}} \frac{\mathscr{C}(\ell, D; \sigma_0)}{\sigma_0^D} \frac{\prod_{\ell=1}^{L-1}(\theta + \ell\,\sigma)}{(\theta_0+1)_{\ell-1}} \frac{\mathscr{C}(q_{r,\cdot,\cdot}, L; \sigma)}{\sigma^L}, \end{aligned}$$

where $\rho(J)$ is the space of the partitions of $[J]$.

The distribution of the overall number of species D in Theorem 5.3.3 is quite involved. However, from such analytical formula we can derive a simple algorithm to sample from it after a variables augmentation.

From the composition structure points out in Theorem 5.3.3 we can also study the asymptotic behavior of the number of species as the sample size n diverges, which boils down to a simple analytical form. From now on, for an arbitrary function $f(n)$, we write $Y_n \asymp f(n)$ if the limit of $Y_n/f(n)$ as n diverges is almost surely a positive and finite random variable.

Theorem 5.3.4. *If the data matrix $\boldsymbol{X}$ is drawn from $(P_1, \ldots, P_J) \sim$ HHPYP and D is the overall number of distinct species in J populations of sample sizes $I_1 = \ldots = I_J = I = n/J$. Then $D \asymp n^{\sigma\sigma_0}$ as $n \to \infty$.*

Notice that the HHPYP can be used also to discover the number of heterogeneous subpopulations R as the number of populations J grows. From (5.11) we have $R \asymp J^{\alpha}$, as $j \to \infty$. That is the number of heterogeneous subpopulations follows a polynomial growth under model (5.3).

5.4 Marginal Gibbs sampler and predictive inference

Posterior inference can be efficiently performed thanks to the marginal Gibbs sampler described in the following section. The full conditionals for the augmented variables S_j and $\boldsymbol{T}_j$ have indeed a nice ratio expression, which is recovered exploiting Bayes theorem and the fact that, with such variable augmentation, the pEPPF admits a product form that simplifies between the numerator and the denominator. This results allow for interpretable and computationally tractable inference for all quantities of interest. These include the posterior distribution of the tables $\boldsymbol{T}$ and, more importantly, the posterior distribution of the vector of cluster assignments $\boldsymbol{S}$ and the predictive distribution of future observations. Such quantities can be used to perform population homogeneity testing, and, for instance, to estimate the number of new species that are expected to be observed in an additional sample of $\boldsymbol{m} = (m_1, \ldots, m_J)$ observations.

5.4.1 Gibbs sampler

The proposed Gibbs sampler follows by extending the marginal Gibbs sampler for NDP mixture models in Zuanetti et al. (2018) to the species sampling framework presented in this chapter. The main idea is that, after having set an initial configuration for the augmented variables $\boldsymbol{S}$ and $\boldsymbol{T}$, at each iteration one first updates the table assigment $T_{j,i}$ for each individual, and then updates the population cluster membership indicators S_j, $j = 1, \ldots, J$, via a Metropolis-Hastings within Gibbs step. Due to the fact that $\boldsymbol{T}_j$ must

be coherent with S_j, the update of S_j is done jointly with an update of $\boldsymbol{T}_j$. The proposal distribution of the Metropolis step is such that it is easy to sample from and allows a fast evaluation of the acceptance probability. Performing homogeneity testing will then be immediate, as it will be sufficient to count the fraction of times two populations are clustered together out of the total number of iterations. In particular, the Gibbs sampler to perform posterior inference on the latent variables $\boldsymbol{S}$ and $\boldsymbol{T}$ is reported below.

(0) At $t = 0$ start from an initial configuration $\boldsymbol{S}$ and $\boldsymbol{T}$.

(1) At iteration $t \geq 1$

(1.a) With $X_{j,i} = X_d^*$ sample latent variables $T_{j,i}$, for $i = 1, \ldots, I_j$ and $j = 1, \ldots, J$ from

$$\text{pr}(T_{j,i} = t \mid \boldsymbol{T}^{-(j,i)}, \boldsymbol{X}, \boldsymbol{S}) \propto \begin{cases} q_{r,t,d}^{-(ji)} - \sigma & \text{if } t = T^*_{r,d,l} \text{ for } l = 1, \ldots, \ell_{r,d}^{-(ji)}, \\ \frac{\omega_d^{-(ji)}}{\ell_{\cdot,\cdot}^{-(ji)} + \theta_0}(\theta + \ell_{r,\cdot}^{-(ji)}\sigma) & \text{if } t = \text{“new”}; \end{cases} \tag{5.12}$$

where $\omega_d^{-(ji)} = \ell_{\cdot,d}^{-(ji)} - \sigma_0$ if $\ell_{\cdot,d}^{-(ji)} > 0$ otherwise $\omega_d^{-(ji)} = 1$.

(1.b) When updating S_j, we will have to update the $\boldsymbol{T}_j$. This is done via the following efficient Metropolis-Hastings within Gibbs step. Call $Y = (S_j, \boldsymbol{T}_j)$ the vector of the current values for the jth population cluster assignment and the table assignments in there, the proposed new values $Y' = (S_j', \boldsymbol{T}_j')$ are sampled by the proposal distribution $q(\cdot \mid \cdot)$, which is defined hierarchically exploiting the results for the importance sampling density in (Maceachern et al., 1999):

$$q(Y' \mid Y) = p(S_j' \mid \boldsymbol{S}_{-j}) \prod_{i=1}^{I_j} p(T_{j,i}' \mid \boldsymbol{T}_{-j}, T_{j,1}', \ldots, T_{j,i-1}', S_j', X_{j,i});$$

where $p(S_j' \mid \boldsymbol{S}_{-j})$ is defined as in (5.8) with $(j+)$ replaced by (j) and, similarly, $p(T_{j,i}' \mid \boldsymbol{T}_{-j}, T_{j,1}', \ldots, T_{j,i-1}', S_j', X_{j,i})$ as in (5.12) replacing (ji) with $\{(j,1), \ldots, (j,i)\}$.

The proposed state Y' is then accepted with probability $\min(1, A')$, where $A' = \frac{p(Y' \mid \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, \boldsymbol{X}) q(Y \mid Y')}{p(Y \mid \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, \boldsymbol{X}) q(Y' \mid Y)}$. Since the full conditional of Y can be expressed as

$$\begin{aligned} p(S_j, \boldsymbol{T}_j \mid \boldsymbol{X}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}) &= \frac{p(S_j, \boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{S}_{-j}, \boldsymbol{T}_{-j})}{p(\boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j})} \propto \\ &\propto p(S_j \mid \boldsymbol{S}_{-j}) p(\boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, S_j), \end{aligned} \tag{5.13}$$

we have

$$A' = \frac{p(\boldsymbol{T}_j', \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, S_j') p(\boldsymbol{T}_j \mid \boldsymbol{X}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, S_j)}{p(\boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, S_j) p(\boldsymbol{T}_j' \mid \boldsymbol{X}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, S_j')},$$

where the conditional distribution for $(\boldsymbol{T}_j, \boldsymbol{X}_j)$ has the form $p(\boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}) =$

$\prod_{i=1}^{I_j} p(\boldsymbol{T}_{j,i}, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, T_{j,1}, \ldots, T_{j,i-1}\boldsymbol{S})$, with

$$\text{pr}(T_{j,i} = t, X_{j,i} = x \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}) =$$
$$\begin{cases} \frac{q_{r,t,d}^{-(ji,\ldots,jI_j)} - \sigma}{\theta + q_{r,\cdot,\cdot}^{-(ji,\ldots,jI_j)}} & \text{if } t = T^*_{r,d,l} \text{ and } x = X^*_d \\ \frac{\omega_d^{-(ji,\ldots,jI_j)}}{\theta_0 + \ell_{\cdot,\cdot}^{-(ji,\ldots,jI_j)}} \frac{\theta + \ell_{r,\cdot}^{-(ji,\ldots,jI_j)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(ji,\ldots,jI_j)}} & \text{if } t = \text{“new”} \text{ and } x = X^*_d \\ \frac{\theta_0 + D^{-(ji,\ldots,jI_j)}\sigma_0}{\theta_0 + \ell_{\cdot,\cdot}^{-(ji,\ldots,jI_j)}} \frac{\theta + \ell_{r,\cdot}^{-(ji,\ldots,jI_j)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(ji,\ldots,jI_j)}} & \text{if } t = \text{“new”} \text{ and } x = \text{“new”}. \end{cases}$$

Finally notice that the denominator in (5.13) is the same for Y and Y' and thus it cancels out when computing A' in the acceptance probability.

5.4.2 Predictive distribution

Consider now the case where we want to make inference about an additional sample of $\boldsymbol{m} = (m_1, \ldots, m_J)$ new observations, where m_j is the number of new observations in population j, for $j = 1, \ldots, J$. Le us denote $\boldsymbol{X}^{\text{new}} = \{X^{\text{new}}_{j,i} : \ j = 1, \ldots, J, \ i = 1, \ldots, m_j\}$ the values of such new observations and $\boldsymbol{T}^{\text{new}} = \{T^{\text{new}}_{j,i} : \ j = 1, \ldots, J, \ i = 1, \ldots, m_j\}$ the latent tables allocations in the HCRF metaphor.

The following urn scheme allows obtain sample $(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}})$ exploiting the output of the Gibbs sampler described in the previous section, since the sample can be obtained sequentially, exploiting the fact that

$$p(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}} \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X}) = \prod_{j=1}^{J} \prod_{i=1}^{m_j} p(X^{\text{new}}_{j,i}, T^{\text{new}}_{j,i} \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{X}^{\text{new}-(ji+)}, \boldsymbol{T}^{\text{new}-(ji+)}),$$

where

$$\text{pr}(X^{\text{new}}_{j,i} = x, T^{\text{new}}_{j,i} = t \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{X}^{\text{new}-(ji+)}, \boldsymbol{T}^{\text{new}-(ji+)}) =$$
$$\begin{cases} \frac{\theta_0 + D^{-(ji+)}\sigma_0}{\theta_0 + \ell_{\cdot,\cdot}^{-(ji+)}} \frac{\theta + \ell_{r,\cdot}^{-(ji+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(ji+)}} \text{ if } x = \text{“new”} \text{ and } t = \text{“new”} \\ \frac{\omega_d^{-(ji+)}}{\theta_0 + \ell_{\cdot,\cdot}^{-(ji+)}} \frac{\theta + \ell_{r,\cdot}^{-(ji+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(ji+)}} \text{ if } x = X^{*-(ji+)}_d \text{ and } t = \text{“new”} \text{ for } d = 1, \ldots, D^{-(ji+)} \\ \frac{q_{r,t,d}^{-(ji+)} - \sigma}{\theta + q_{r,\cdot,\cdot}^{-(ji+)}} \text{ if } x = X^{*-(ji+)}_d \text{ and } t = T^*_{r,d,l} \text{ for } l = 1, \ldots, \ell_{r,d}^{-(ji+)}, \ d = 1, \ldots, D^{-(ji+)}, \end{cases}$$

being $(ji+) = \{(jl) : l \geq i\} \cup \{(kl) : k \geq j\}$ the index set associated to the future random variables not sampled yet.

Thus, for each configuration $(\boldsymbol{S}, \boldsymbol{T})$ generated in the Gibbs sampler presented in Section 5.4.1, one can obtain a sample from $p(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}} \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X})$, so that, after the burn-in period, samples from $p(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}} \mid \boldsymbol{X})$ are obtained.

5.A Appendix: Proofs

5.A.1 Proof of Equations (5.4) and (5.5)

Notice that $P_j \overset{\mathrm{d}}{=} P_1^*$.

$$\begin{aligned}\mathbb{E}[P_j(A)] &= \mathbb{E}[P_1^*(A)] = H(A) \text{ since } P_1^* \text{ is a species sampling model}\\ \mathrm{Var}[P_j(A)] &= \mathrm{Var}[P_1^*(A)]\end{aligned}$$

We also know that $\mathrm{Var}[P_0^*(A)] = H(A)[1-H(A)]\dfrac{1-\sigma_0}{\theta_0+1}$ and $\mathrm{Var}[P_1^*(A)] = \dfrac{H(A)[1-H(A)]}{\theta_0+1}\left[(1-\sigma_0)+(\theta_0+\sigma_0)\dfrac{1-\sigma}{\theta+1}\right]$, see Camerlenghi et al. (2019b).

Moreover $\mathbb{E}[P_1^*(A)P_2^*(A)] = \mathbb{E}[\mathbb{E}[P_1^*(A) \mid P_0^*]\mathbb{E}[P_2^*(A) \mid P_0^*]] = \mathbb{E}[P_0^*(A)^2]$ and $\mathrm{pr}(P_j = P_{j'}) = \dfrac{1-\alpha}{\gamma+1}$ for $j \neq j'$. Thus,

$$\begin{aligned}\mathbb{E}[P_j(A)P_{j'}(A)] &= \mathbb{E}[P_1(A)P_2(A) \mid P_1 = P_2]\mathrm{pr}(P_1 = P_2) + \mathbb{E}[P_1(A)P_2(A) \mid P_1 \neq P_2]\mathrm{pr}(P_1 \neq P_2)\\ &= \frac{1-\alpha}{\gamma+1}\mathbb{E}[P_1^*(A)^2] + \frac{\gamma+\alpha}{\gamma+1}\mathbb{E}[P_1^*(A)P_2^*(A)]\\ &= \frac{1-\alpha}{\gamma+1}\mathbb{E}[P_1^*(A)^2] + \frac{\gamma+\alpha}{\gamma+1}\mathbb{E}[P_0^*(A)^2].\end{aligned}$$

From this we obtain

$$\mathrm{Cov}[P_j(A), P_{j'}(A)] = \mathbb{E}[P_j(A)P_{j'}(A)] - H(A)^2 = \frac{1-\alpha}{\gamma+1}\mathrm{Var}[P_1^*(A)^2] + \frac{\gamma+\alpha}{\gamma+1}\mathrm{Var}[P_0^*(A)^2]$$

and

$$\begin{aligned}\mathrm{Cor}[P_j(A), P_{j'}(A)] &= \frac{\mathrm{Cov}[P_j(A), P_{j'}(A)]}{\mathrm{Var}[P_1^*(A)]} = \frac{1-\alpha}{\gamma+1} + \frac{\gamma+\alpha}{\gamma+1}\frac{\mathrm{Var}[P_0^*(A)]}{\mathrm{Var}[P_1^*(A)]}\\ &= \frac{1-\alpha}{\gamma+1} + \frac{\gamma+\alpha}{\gamma+1}\frac{1-\sigma_0}{(1-\sigma_0)+(\theta_0+\sigma_0)\dfrac{1-\sigma}{\theta+1}}\\ &= \frac{1-\alpha+\dfrac{(\alpha+\gamma)(-1+\sigma_0)(1+\theta)}{-1+(-1+\sigma_0)\theta-\theta_0+\sigma(\sigma_0+\theta_0)}}{1+\gamma}\end{aligned}$$

□

5.A.2 Proof of Equation (5.6)

Notice that $X_{j,i} \overset{\mathrm{d}}{=} X_l^*$. Thus,

$$\begin{aligned}\mathrm{Cov}[X_{j,i}, X_{j',i'}] &= \mathbb{E}[\mathrm{Cov}(X_{j,i} = X_{j',i'} \mid \mathbf{1}(X_{j,i} = X_{j',i'}))]\mathrm{pr}(X_{j,i} = X_{j',i'}) + 0\\ &= \mathrm{pr}(X_{j,i} = X_{j',i'})\mathrm{Var}(X_l^*)\end{aligned}$$

Therefore $\mathrm{Cor}[X_{j,i}, X_{j',i'}] = \mathrm{pr}(X_{j,i} = X_{j',i'})$, where

$$\begin{aligned}\mathrm{pr}(X_{j,i'} = X_{j,i}) &= \mathrm{pr}(X_{j,i'} = X_{j,i} \mid T_{j,i} = T_{j,i'})\mathrm{pr}(T_{j,i} = T_{j,i'}) + \\ &\quad \mathrm{pr}(T_{j,i} \neq T_{j,i'})\mathrm{pr}(X_{j,i'} = X_{j,i} \mid T_{j,i} \neq T_{j,i'}) \\ &= \frac{1-\sigma}{\theta+1} + \frac{1-\sigma_0}{\theta_0+1}\frac{\theta+\sigma}{\theta+1}\end{aligned}$$

and, if $j \neq j'$,

$$\begin{aligned}\mathrm{pr}(X_{j,i'} = X_{j,i}) &= \mathrm{pr}(X_{j,i} = X_{j',i'} \mid P_j = P_{j'})\mathrm{pr}(P_j = P_{j'}) + \\ &\quad \mathrm{pr}(X_{j,i} = X_{j',i'} \mid P_j \neq P_{j'})\mathrm{pr}(P_j \neq P_{j'}) \\ &= \left\{\left[\frac{1-\sigma}{\theta+1} + \frac{1-\sigma_0}{\theta_0+1}\frac{\theta+\sigma}{\theta+1}\right](1-\alpha) + \frac{1-\sigma_0}{\theta_0+1}(\gamma+\alpha)\right\}(\gamma+1)^{-1}\end{aligned}$$

□

5.A.3 Proof of Theorem 5.3.1

In order to prove Theorem 5.3.1 we first show that the following lemma holds true.

Lemma 5.A.1. *The random partition structure induced by the samples $\boldsymbol{X}$ drawn from $(P_1, \ldots, P_J) \sim$ HHPYP given a particular partition of distributions $\Psi^{(J)} = \{B_1, \ldots, B_R\}$ is characterized by the pEPPF*

$$\Pi_D^{(n)}\left(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J \mid \Psi^{(J)} = \{B_1, \ldots, B_R\}\right) = \Phi_D^{(n)}\left(q_{1,\cdot,\cdot}, \ldots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0\right),$$

where $\Phi_D^{(n)}(q_{1,\cdot,\cdot}, \ldots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0)$ denotes the pEPPF associated to an R-dimensional $\mathrm{HPYP}(\sigma, \sigma_0, \theta, \theta_0; H)$.

Indeed,

$$\begin{aligned}&\Pi_D^{(n)}\big(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J \mid \Psi^{(J)} = \{B_1, \ldots, B_R\}\big) = \\ &= \mathbb{E}\left[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} P_1^{n_{1,d}}(\mathrm{d}x_d) \ldots P_J^{n_{J,d}}(\mathrm{d}x_d) \;\middle|\; \Psi^{(J)} = \{B_1, \ldots, B_R\}\right] = \\ &= \mathbb{E}\left[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} P_1^{*q_{1,\cdot,d}}(\mathrm{d}x_d) \ldots P_R^{*q_{r,\cdot,d}}(\mathrm{d}x_d)\right] = \Phi_D^{(n)}(\boldsymbol{n}_1^*, \ldots, \boldsymbol{n}_R^*; \sigma, \theta, \sigma_0, \theta_0),\end{aligned}$$

where $\mathbb{X}_*^D = \mathbb{X}^D \setminus \{\boldsymbol{x} : x_i = x_j \text{ for some } i \neq j\}$ and $(P_1^*, \ldots, P_R^*) \sim \mathrm{HPYP}(\sigma, \sigma_0, \theta, \theta_0; H)$. Moreover, notice that the R unique values between $(P_1, \ldots, P_J)$ are not necessary the first $(P_1^*, \ldots, P_R^*)$ but since $(P_k^*)_{k \geq 1}$ are exchangeable the third equality holds. Therefore, applying Lemma 5.A.1

$$\begin{aligned}\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) &= \sum \mathrm{pr}(\Psi^{(J)} = \{B_1, \ldots, B_D\})\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J \mid \Psi^{(J)} = \{B_1, \ldots, B_D\}) \\ &= \sum \phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma)\Phi_D^{(n)}(q_{1,\cdot,\cdot}, \ldots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0)\end{aligned}$$

□

5.A.4 Proof of Proposition 5.3.2

In order to derive the posterior probability of degeneracy we rewrite the marginal likelihood as

$$p(\boldsymbol{X}) = \Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) \prod_{d=1}^{D} H(\mathrm{d}\boldsymbol{X}_d^*),$$

where $\{\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_D^*\}$ are the D unique values between $\boldsymbol{X}$ and $\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2)$ is the pEPPF associated to the proposed model 5.9, that is

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \mathrm{pr}(P_1 = P_2)\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \mathrm{pr}(P_1 \neq P_2)\Phi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0),$$

Finally we prove the proposition by applying Bayes theorem

$$\begin{aligned}\mathrm{pr}(P_1 = P_2 \mid \boldsymbol{X}) &= \frac{\mathrm{pr}(P_1 = P_2)p(\boldsymbol{X} \mid P_1 = P_2)}{p(\boldsymbol{X})} \\ &= \frac{(1-\alpha)\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}{(1-\alpha)\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0) + (\alpha + \gamma)\,\Phi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}.\end{aligned}$$

□

5.A.5 Proof of Theorem 5.3.3

Notice that applying Lemma 5.A.1 and Theorem 6 in (Camerlenghi et al., 2019b) we have that

$$p(D \mid \Psi^{(J)} = \{B_1, \ldots, B_R\}) = \sum_{L=D}^{n} \mathrm{pr}(D_{0,L} = D)\,\mathrm{pr}\left(\sum_{j=1}^{J} L_r = L\right).$$

Then marginalizing out the population partition $\Psi^{(J)}$ we have

$$p(D) = \sum_{\boldsymbol{B} \in \rho(J)} \phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma) \sum_{L=D}^{n} \mathrm{pr}(D_{0,L} = D)\,\mathrm{pr}\left(\sum_{j=1}^{J} L_r = L\right).$$

□

5.A.6 Proof of Theorem 5.3.4

Let $T(\boldsymbol{n}) \overset{\mathrm{d}}{=} \sum_{r=1}^{R} L_r \leq D$, representing the number of tables in the franchise. The conditional independence arising from the hierarchical specification of the model (5.3) entails that $D = D_{0,T(\boldsymbol{n})}$ almost surely. Moreover, by the asymptotic of the number of species in the exchangeable case under a Pitman–Yor prior we have that for each $m_r = m_r(\Psi^{(J)}) \in \{0, \ldots, J\}$:

$$\frac{D_{0,I}}{I^{\sigma_0}} \xrightarrow{\text{a.s.}} C_0, \qquad \frac{L_r}{I^{\sigma}} \xrightarrow{\text{a.s.}} C_r m_r^{\sigma},$$

as $I \to \infty$, where C_0 and C_r's are positive and finite random variables. Since $T(\boldsymbol{n}) = \sum_r^R L_{r,m_r I}$

$$\frac{T(\boldsymbol{n})}{I^\sigma} \xrightarrow{\text{a.s.}} \sum_{r=1}^{R} C_r m_r^\sigma = \eta(\Psi^{(J)}),$$

where $\eta = \eta(\Psi^{(J)})$ is a positive finite random variable. Thus,

$$\frac{D_{0,T(\boldsymbol{n})}}{D_{0,\eta I^\sigma}} = \frac{T(\boldsymbol{n})^{\sigma_0}}{(\eta I^\sigma)^{\sigma_0}} \frac{D_{0,T(\boldsymbol{n})}/T(\boldsymbol{n})^{\sigma_0}}{D_{0,\eta I^\sigma}/(\eta I^\sigma)^{\sigma_0}} \xrightarrow{\text{a.s.}} 1.$$

entailing

$$\frac{D_n}{I^{\sigma\sigma_0}} = \frac{D_{0,T(\boldsymbol{n})}}{D_{0,\eta I^\sigma}} \frac{D_{0,\eta I^\sigma}}{(I^\sigma)^{\sigma_0}} \xrightarrow{\text{a.s.}} C_0,$$

as $I \to \infty$. □

Bibliography

Agarwal, R., Ranjan, P., and Chipman, H. (2014). A new Bayesian ensemble of trees approach for land cover classification of satellite imagery. *Canadian Journal of Remote Sensing*, 39:507–520.

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

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679.

Albert, J. H. and Chib, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics*, 57(3):829–836.

Andrieu, C. and Doucet, A. (2002). Particle filtering for partially observed Gaussian state space models. *Journal of the Royal Statistical Society: Series B*, 64(4):827–836.

Arellano-Valle, R. B. and Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33:561–574.

Armagan, A. and Zaretzki, R. L. (2011). A note on mean-field variational approximations in Bayesian probit models. *Computational Statistics & Data Analysis*, 55(1):641–643.

Arnold, B. C. and Beaver, R. J. (2000). Hidden truncation models. *Sankhyā: Series A*, 62:23–35.

Arnold, B. C., Beaver, R. J., Azzalini, A., Balakrishnan, N., Bhaumik, A., Dey, D., Cuadras, C., and Sarabia, J. M. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Test*, 11:7–54.

Atkins, A., Niranjan, M., and Gerding, E. (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, 4(2):120–137.

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

Azzalini, A. and Bacchieri, A. (2010). A prospective combination of phase II and phase III in drug development. *Metron*, 68(3):347–369.

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

Azzalini, A. and Capitanio, A. (2014). *The Skew-normal and Related Families*. Cambridge University Press.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83:715–726.

Battiston, M., Favaro, S., and Teh, Y. W. (2018). Multi-armed bandit for species discovery: a Bayesian nonparametric approach. *Journal of the American Statistical Association*, 113(521):455–466.

Beraha, M., Guglielmi, A., and Quintana, F. A. (2020). The semi-hierarchical Dirichlet Process and its application to clustering homogeneous distributions. *arXiv*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Börsch-Supan, A. and Hajivassiliou, V. A. (1993). Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics*, 58(3):347–368.

Botev, Z. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B*, 79(1):125–148.

Brown, L. D., Cai, T. T., Low, M. G., and Zhang, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *The Annals of Statistics*, 30(3):688–707.

Burgette, L. F. and Nordheim, E. V. (2012). The trace restriction: An alternative identification strategy for the bayesian multinomial probit model. *Journal of Business & Economic Statistics*, 30(3):404–410.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodríguez, A. (2019a). Latent nested nonparametric priors (with discussion). *Bayesian Analysis*, 14(4):1303–1356.

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019b). Distribution theory for hierarchical processes. *The Annals of Statistics*, 47(1):67–92.

Camerlenghi, F., Lijoi, A., and Prünster, I. (2017). Bayesian prediction with multiple-samples information. *Journal of Multivariate Analysis*, 156:18–28.

Cao, J., Genton, M., Keyes, D., and Turkiyyah, G. (2019). Hierarchical–block conditioning approximations for high-dimensional multivariate normal probabilities. *Statistics and Computing*, 29:585–598.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992). A Monte Carlo approach to non-normal and nonlinear state-space modeling. *Journal of the American Statistical Association*, 87(418):493–500.

Charalambides, C. A. (2002). *Enumerative Combinatorics*. Chapman and Hall/CRC.

Chen, Z. and Kuo, L. (2002). Discrete choice models based on the scale mixture of multivariate normal distributions. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 192–213.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85:347–361.

Chipman, H. A., George, E. I., and 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(2):275–288.

Chopin, N. and Ridgway, J. (2017). Leave Pima indians alone: Binary regression as a benchmark for Bayesian computation. *Statistical Science*, 32:64–87.

Christensen, J. and Ma, L. (2020). A Bayesian hierarchical model for related densities by using Pólya trees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):127–153.

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

Craig-Schapiro, R., Kuhn, M., Xiong, C., Pickering, E. H., Liu, J., Misko, T. P., Perrin, R. J., Bales, K. R., Soares, H., and Fagan, A. M. (2011). Multiplexed immunoassay panel identifies novel CSF biomarkers for Alzheimer's disease diagnosis and prognosis. *PloS one*, 6(4):e18850.

Daganzo, C. (2014). *Multinomial probit: the theory and its application to demand forecasting*. Elsevier.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229.

Doucet, A., De Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208.

Doucet, A. and Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: fifteen years later. *Handbook of Nonlinear Filtering*, 12(3):656–704.

Dow, J. K. and Endersby, J. W. (2004). Multinomial probit and multinomial logit: a comparison of choice models for voting research. *Electoral studies*, 23(1):107–122.

Durante, D. (2019). Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106(4):765–779.

Durante, D. and Rigon, T. (2019). Conditionally conjugate mean-field variational Bayes for logistic models. *Statistical Science*, 34(3):472–485.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Favaro, S., Lijoi, A., Mena, R. H., and Prünster, I. (2009). Bayesian non-parametric inference for species variety with a two-parameter Poisson-Dirichlet process prior. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):993–1008.

Favaro, S., Lijoi, A., and Prünster, I. (2012). A new estimator of the discovery probability. *Biometrics*, 68(4):1188–1196.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

de Finetti, Bruno (1938). Sur la condition d'equivalence partielle. *Actualitès Scientifiques et Industrielles*, 739:5–18.

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

Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., Riebler, A., et al. (2018). Intuitive joint priors for variance parameters. *Bayesian Analysis*.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Genton, M. G., Keyes, D. E., and Turkiyyah, G. (2018). Hierarchical decompositions for the computation of high-dimensional multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 27(2):268–277.

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

Geweke, J., Keane, M., and Runkle, D. (1994). Alternative computational approaches to inference in the multinomial probit model. *The review of economics and statistics*, pages 609–632.

Girolami, M. and Rogers, S. (2006). Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18:1790–1817.

Girolami, M. and Zhong, M. (2007). Data integration for classification problems employing Gaussian process priors. In *Advances in Neural Information Processing Systems*, volume 20, pages 465–472.

Goldwater, S., Johnson, M., and Griffiths, T. L. (2006). Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems*, pages 459–466.

González-Farías, G., Domínguez-Molina, A., and Gupta, A. K. (2004). Additive properties of skew normal random vectors. *Journal of Statistical Planning and Inference*, 126:521–534.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Good, I. J. and Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F-radar and Signal Processing*, 140(2):107–113.

Green, P. J., Łatuszyński, K., Pereyra, M., and Robert, C. P. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Gupta, A. K., Aziz, M. A., and 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-Farías, G., and Domínguez-Molina, J. A. (2004). A multivariate skew normal distribution. *Journal of Multivariate Analysis*, 89:181–190.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

Hausman, J. A. and Wise, D. A. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica: Journal of the econometric society*, pages 403–426.

Hoffman, M. D. and 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. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1:145–168.

Horrace, W. C. (2005). Some results on the multivariate truncated normal distribution. *Journal of Multivariate Analysis*, 94(1):209–221.

Imai, K. and Van Dyk, D. A. (2005). A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of econometrics*, 124(2):311–334.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Johndrow, J., Dunson, D., and Lum, K. (2013). Diagonal orthant multinomial probit models. In *Artificial Intelligence and Statistics*, pages 29–38.

Johndrow, J. E., Smith, A., Pillai, N., and Dunson, D. B. (2019). MCMC for imbalanced categorical data. *Journal of the American Statistical Association*, 114(527):1394–1403.

Julier, S. J. and Uhlmann, J. K. (1997). New extension of the Kalman filter to nonlinear systems. In *Proceedings SPIE 3068, Signal Processing, Sensor Fusion, and Target Recognition*, pages 182–194.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.

Kara, Y., Boyacioglu, M. A., and Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert Systems with Applications*, 38(5):5311–5319.

Keane, M. P. and Wolpin, K. I. (2009). Empirical applications of discrete choice dynamic programming models. *Review of Economic Dynamics*, 12(1):1–22.

Kim, K.-j. and Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications*, 19(2):125–132.

Kindo, B., Wang, H., and Peña, E. (2016). Multinomial probit Bayesian additive regression trees. *Stat*, 5:119–131.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25.

Knowles, D. and Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems*, volume 24, pages 1701–1709.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

Kuss, M. and Rasmussen, C. E. (2005). Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786.

Lijoi, A., Mena, R. H., and Prünster, I. (2008). A Bayesian nonparametric approach for comparing clustering structures in EST libraries. *Journal of Computational Biology*, 15(10):1315–1327.

Lijoi, A., Muliere, P., Prünster, I., and Taddei, F. (2016). Innovation, growth and aggregate volatility from a bayesian nonparametric perspective. *Electronic Journal of Statistics*, 10(2):2179–2203.

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044.

Liu, X., Daniels, M. J., and Marcus, B. (2009). Joint models for the association of longitudinal binary and continuous processes with application to a smoking cessation trial. *Journal of the American Statistical Association*, 104(486):429–438.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. CRC Press.

Maceachern, S. N., Clyde, M., and Liu, J. S. (1999). Sequential importance sampling for nonparametric Bayes models: the next generation. *Canadian Journal of Statistics*, 27(2):251–267.

Maddala, G. S. (1986). *Limited-dependent and qualitative variables in econometrics*. Number 3. Cambridge university press.

McCulloch, R. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1-2):207–240.

McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000). A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of econometrics*, 99(1):173–193.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica: Journal of the Econometric Society*, 57:995–1026.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., and Bartoli, A. (2016). Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE transactions on medical imaging*, 35(9):2051–2063.

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of Uncertainty in Artificial Intelligence*, volume 17, pages 362–369.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251.

Natarajan, R., McCulloch, C. E., and Kiefer, N. M. (2000). A Monte Carlo EM method for estimating multinomial probit models. *Computational Statistics & Data Analysis*, 34(1):33–50.

Nishimura, A. and Suchard, M. A. (2018). Prior-preconditioned conjugate gradient for accelerated gibbs sampling in "large n & large p" sparse bayesian logistic regression models. *arXiv preprint arXiv:1810.12437*.

Nobile, A. (1998). A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statistics and Computing*, 8(3):229–242.

Pakman, A. and Paninski, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models with R*. Springer.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. *Statistics, Probability and Game Theory. Papers in honor of David Blackwell*, pages 245–267.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108:1339–1349.

Qin, Q. and Hobert, J. P. (2019). Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression. *The Annals of Statistics*, 47(4):2320–2347.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Reiß, M. (2008). Asymptotic equivalence for nonparametric regression with multivariate and random design. *The Annals of Statistics*, 36(4):1957–1982.

Riihimäki, J., Jylänki, P., and Vehtari, A. (2013). Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *Journal of Machine Learning Research*, 14:75–109.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367.

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6:145–178.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):483–1131.

Rogers, S. and Girolami, M. (2007). Multi-class semi-supervised learning with the ϵ-truncated multinomial probit gaussian process. In *Journal of Machine Learning Research, Workshop & Proceedings*, volume 1, pages 17–32.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika*, 81(1):115–131.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28.

Soriano, J. and Ma, L. (2019). Mixture modeling on related samples by ψ-stick breaking and kernel perturbation. *Bayesian Analysis*, 14(1):161–180.

Soyer, R. and 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: Journal of the Econometric Society*, 60:943–952.

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, pages 985–992, Morristown, NJ, USA. Association for Computational Linguistics.

Teh, Y. W. and Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian nonparametrics*, chapter 5, pages 158–207. Cambridge University Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Tutz, G. (1991). Sequential models in categorical regression. *Computational Statistics & Data Analysis*, 11(3):275–295.

Uhlmann, J. K. (1992). Algorithms for multiple-target tracking. *American Scientist*, 80(2):128–141.

Villani, C. (2008). *Optimal Transport: Old and New*. Springer Science & Business Media.

West, M. and Harrison, J. (2006). *Bayesian Forecasting and Dynamic Models*. Springer Science & Business Media.

Zhang, X., Boscardin, W. J., and Belin, T. R. (2006). Sampling correlation matrices in bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, 15(4):880–896.

Zuanetti, D. A., Müller, P., Zhu, Y., Yang, S., and Ji, Y. (2018). Clustering distributions with the marginalized nested Dirichlet process. *Biometrics*, 74(2):584–594.