Biometrika (2024), **103**, 1, *p*. 1
Printed in Great Britain Advance Access publication on 31 July 2023

Multivariate Gaussian cumulative distribution functions as the marginal likelihood of their dual Bayesian probit models

By AUGUSTO FASANO

Dept. of Economics, Social Studies, Applied Mathematics and Statistics, University of Torino, Corso Unione Sovietica 218 bis, 10134 Torino, Italy, and Collegio Carlo Alberto
augusto.fasano@unito.it

and FRANCESCO DENTI

Dept. of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy
francesco.denti@unipd.it

Summary

The computation of multivariate Gaussian cumulative distribution functions is a key step in many statistical procedures, often representing a crucial computational bottleneck. Over the last decades, efficient algorithms have been proposed to address this problem, mainly leveraging on Monte Carlo solutions. This article highlights a connection between the multivariate Gaussian cumulative distribution function and the marginal likelihood of a tailored dual Bayesian probit model. Consequently, any method that approximates such a marginal likelihood can be used to estimate the quantity of interest. In this work, we focus on the approximation provided by the expectation propagation algorithm. Its empirical accuracy and polynomial computational cost make it an appealing choice, especially for tail probabilities, even if theoretical guarantees are currently limited. Its efficiency, accuracy, and stability are shown for multiple correlation matrices and integration limits, highlighting a series of advantages over state-of-the-art alternatives.

Some key words: Approximate inference; Cumulative distribution function; Expectation propagation; Multivariate Gaussian distribution; Probit model.

1. Introduction

The computation of multivariate Gaussian cumulative distribution functions is a problem encountered in many statistical applications, often representing the computational bottleneck in model fitting, parameter estimation, and prediction. Let us consider, without loss of generality, an m-dimensional multivariate Gaussian random variable with mean 0 and covariance matrix Σ. Then, for $u = (u_1, \dots, u_m)^\top \in \mathbb{R}^m$, its cumulative distribution function is defined as

$$\Phi_m(u; \Sigma) = \int_{C_u} \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} \exp\left\{-\frac{1}{2} w^\top \Sigma^{-1} w\right\} dw, \tag{1}$$

where $C_u = \{w = (w_1, \dots, w_m)^\top \in \mathbb{R}^m : w_i \le u_i, \ \forall i = 1, \dots, m\}$. Such functions usually appear as a consequence of partially-observed Gaussian latent variables in various models. Popular examples include the Bayesian probit (Albert & Chib, 1993; Durante, 2019), the multivariate (Ochi & Prentice, 1984) and multinomial (Hausman & Wise, 1978; Stern, 1992; Connors et al., 2014; Fasano & Durante, 2022; Ding et al., 2024) probit, and formulations that employ skew-

normal densities (Genton, 2004; Arellano-Valle & Azzalini, 2006; Azzalini, 2014) to account for skewness in the observations and, possibly, in the model parameters under the Bayesian setting. See Anceschi et al. (2023a) for a broad class of Bayesian models whose posterior distribution belongs to the unified skew-normal family (Arellano-Valle & Azzalini, 2006), thus involving multivariate Gaussian cumulative distribution functions in their posterior density. Sampling algorithms for the computation of $\Phi_m(u;\Sigma)$ typically leverage on the separation-of-variables method developed by Genz (1992), where the integration region is converted into the unit hypercube thanks to a change of variable. However, its high computational cost motivated further research to devise more scalable and accurate alternatives. Trinh & Genz (2015) considered both univariate and bivariate reorderings of variables, observing that estimation accuracy can be effectively improved by reordering the variables so that outermost ones have the smallest expected values. Botev (2017) improved the accuracy over the separation-of-variables method by introducing a minimax tilting technique. Although it has an asymptotically vanishing relative error, such an approach may become computationally impractical for dimensions above a few hundred. Other proposals focusing on orthant probabilities include the sequential Monte Carlo approach by Ridgway (2016), the two-step method by Azzimonti & Ginsbourger (2018), and previous approaches exploiting the specific orthant structure (Miwa et al., 2003; Craig, 2008). See also Genz & Bretz (2009). To obtain more scalable integral estimates, Cao et al. (2021) proposed a procedure that combines the separation-of-variables method with the tile-low-rank representation (Weisbecker, 2013; Mary, 2017; Akbudak et al., 2017) of the covariance matrix and the block variable reordering from Cao et al. (2019), improving over the hierarchical quasi Monte Carlo method by Genton et al. (2018). This representation leads to massive computational advantages, increasing the dimensions of the problems that can be effectively tackled, as the method is feasible for dimensions of m of the order of tens of thousands. However, the maximal computational gains are obtained when the covariance structure exhibits low-rank structures. When this is not the case, the tile-low-rank method may not be applicable, leaving the computation of multivariate Gaussian cumulative distribution functions with a generic covariance matrix an open issue.

In the present contribution, we address the problem by showing that any Gaussian cumulative distribution function (1) can be characterized as the marginal likelihood of a *dual* Bayesian probit model. As a consequence, any method that gives an approximation of the marginal likelihood can be employed to estimate $\Phi_m(u;\Sigma)$. Popular options for this task are available both among sampling methods (e.g., sequential Monte Carlo, Chopin & Papaspiliopoulos, 2020), and deterministic approximations (e.g., variational inference, Blei et al., 2017, or expectation propagation Minka, 2001a,b; Vehtari et al., 2020). Motivated by the excellent empirical performance of the expectation propagation algorithm in various applications (Chopin & Ridgway, 2017; Braunstein et al., 2017; Zhang et al., 2019; Vehtari et al., 2020; Zhou et al., 2023; Anceschi et al., 2023a), we exploit such an algorithm to approximate $\Phi_m(u;\Sigma)$. This method avoids the computational issues associated with sampling and is shown to be accurate and stable for generic covariance matrices, although theoretical guarantees are currently confined to the large data limit (Dehaene & Barthelmé, 2015, 2018). As shown in the experiments, expectation propagation may be the preferable option for computing tail probabilities, where sampling methods may encounter underflow issues, or in high-dimensional settings, due to its computational advantages. Such an approach is also motivated by the promising results obtained by Cunningham et al. (2011), who used expectation propagation for the computation of Gaussian probabilities, although with a different perspective from the one presented here. In their approximation, the multivariate Gaussian probability of interest is seen as the normalizing constant of a truncated multivariate Gaussian density function, and expectation propagation is applied by considering the unconstrained Gaussian density as the prior distribution and the indicator functions of the truncation constraints as likelihood terms.

2. Gaussian cumulative distribution functions as marginal likelihoods of dual Bayesian probit models

Consider a Bayesian probit model with n observations, p explanatory variables, and Gaussian prior with mean vector ξ and covariance matrix Ω for the coefficients. Let $y = (y_1, \dots, y_n)^\top \in \{0,1\}^n$ denote the observation vector and $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times p}$ be the design matrix with generic row $x_i^\top$ given by the covariate vector for observation i. Then, this model can be defined as

$$y_i \mid \beta \overset{ind}{\sim} \text{Bernoulli}\left\{\Phi_1(x_i^\top \beta; 1)\right\}, \quad i = 1, \dots, n, \quad \beta \sim \mathcal{N}_p(\xi, \Omega). \tag{2}$$

The crucial point of our contribution is recognizing that any function of the form (1) can be rewritten as the marginal likelihood of a dual Bayesian probit model with $n = p = m$. This fundamental analogy is guaranteed by the following proposition.

Proposition 1. *Given any positive-definite $m \times m$ covariance matrix Σ, call λ_m its smallest eigenvalue and define $\tilde{\Sigma} = \Sigma - \epsilon\lambda_m I_m$, where $\epsilon \in (0,1)$. Consider now any factorization of $\tilde{\Sigma}$ of the form $\tilde{\Sigma} = P\tilde{\Lambda}P^\top$, where P and $\tilde{\Lambda}$ are $m \times m$ matrices with the former invertible and the latter symmetric and positive definite. Then, for any $u \in \mathbb{R}^m$, $\Phi_m(u; \Sigma)$ equals the marginal likelihood of a dual Bayesian probit model (2) in which $n = p = m$, $y_i = 1$ for all $i = 1, \dots, n$, and*

$$\xi = (\epsilon\lambda_m)^{-1/2} P^{-1} u, \quad \Omega = (\epsilon\lambda_m)^{-1} \tilde{\Lambda}, \quad X = P.$$

The proof is reported in the Supplementary Material. Since $\tilde{\Sigma}$ in Proposition 1 is still a positive-definite symmetric $m \times m$ matrix, whose eigenvalues simply differ from the ones of Σ by $\epsilon\lambda_m$, multiple factorizations of the form $P\tilde{\Lambda}P^\top$ are possible, like the eigendecomposition, the singular value, and the Cholesky decomposition (Petersen et al., 2008). In the experiments, we considered the eigen- and Cholesky decompositions. In the former, P is an orthogonal matrix whose columns are the eigenvectors of $\tilde{\Sigma}$, and $\tilde{\Lambda} = \text{diag}(\tilde{\lambda}_1, \dots, \tilde{\lambda}_m)$ is a diagonal matrix whose entries are the sorted eigenvalues of $\tilde{\Sigma}$ in decreasing order; thus, the Gaussian prior for β in the dual probit has mean $(\epsilon\lambda_m)^{-1/2} P^\top u$ and independent components with variances $\tilde{\lambda}_i/(\epsilon\lambda_m)$, $i = 1, \dots, m$. In the Cholesky decomposition, P is lower triangular and $\tilde{\Lambda} = I_m$, so that β has a spherical Gaussian prior, with mean $(\epsilon\lambda_m)^{-1/2} P^{-1} u$ and independent components with equal variance $(\epsilon\lambda_m)^{-1}$. The two decompositions are shown to give comparable results in terms of running times and accuracy of the approximation in the Supplementary Material. As a direct implication of Proposition 1, any method that efficiently approximates the marginal likelihood of the Bayesian probit model (2) can be leveraged to compute the Gaussian cumulative distribution function (1). In the remainder of the paper, we concentrate on the approximation given by the expectation propagation algorithm.

3. Expectation propagation for Gaussian cumulative distribution functions

Motivated by the empirical accuracy showed in multiple related applications (Chopin & Ridgway, 2017; Zhang et al., 2019; Fasano et al., 2023; Anceschi et al., 2023a,b, 2024), we rely on expectation propagation to approximate the marginal likelihood in the dual probit model of Proposition 1, obtaining an estimate of $\Phi_m(u; \Sigma)$ in (1). In general, this method consists in an iterative algorithm to obtain an approximation of the posterior distribution, its moments, and, if desired, the marginal likelihood in a Bayesian model with conditionally independent observations having likelihood $p(y_i \mid \beta)$, $i = 1, \dots, n$, and parameter β *a priori* distributed as $p(\beta)$. To this scope, the posterior distribution $p(\beta \mid y) \propto p(\beta) \prod_{i=1}^n p(y_i \mid \beta)$ of the p-dimensional parameter β is approximated with a density $q_{\text{EP}}(\beta)$ that has the same factorization $q_{\text{EP}}(\beta) \propto q_0(\beta) \prod_{i=1}^n q_i(\beta)$. To have a tractable global approximation $q_{\text{EP}}(\beta)$, the factors (or *sites*) $q_i(\beta)$, $i = 0, \dots, n$, are taken from an exponential family kernel whose parameters are updated sequentially imposing

some *moment matching* conditions. Details can be found in Section S.2 of the Supplementary Material. See also Fasano et al. (2023) for detailed derivations for the Bayesian probit model (excluding the computation of the marginal likelihood) and Anceschi et al. (2024) for a complete overview in the case of multiple generalized linear models, including the probit model. The key point for the expectation propagation implementation is the tractability of the *hybrid density* $h_i(\beta) = p(y_i \mid \beta) q_{-i}(\beta)/Z_{h_i}$, where $q_{-i}(\beta) \propto q_{\text{EP}}(\beta)/q_i(\beta)$ is the so-called *cavity density* and $Z_{h_i} = \int p(y_i \mid \beta) q_{-i}(\beta) d\beta$. Indeed, one has to be able to compute the first two moments of $h_i(\beta)$ as well as the normalizing constant Z_{h_i}. In the case of the probit model, this is immediate since $h_i(\beta)$ is the density of a multivariate extended skew-normal (Anceschi et al., 2024; Azzalini, 2014) and all the desired quantities are available in closed form.

Classic expectation propagation implementations for a Bayesian probit model with p-dimensional parameter and n observations have per-iteration cost $O(p^2 n)$ (Chopin & Ridgway, 2017; Anceschi et al., 2024), although different implementations with $O(pn^2)$ per-iteration cost are possible, which tend to be more efficient when both p and n are large and are of the same order (see Algorithm 2 in Anceschi et al., 2024). Since in the dual probit model (2) $n = p = m$, the overall cost to compute $\Phi_m(u; \Sigma)$ via expectation propagation, including preprocessing, has order $O(m^3)$, regardless of the specific implementation considered. This is also in line with the $O(n^3)$ per-iteration cost reported for Gaussian processes classification via expectation propagation for n observations with probit likelihood in Chapter 3.6 of Rasmussen & Williams (2006).

In our studies, we have implemented the expectation propagation routine for the dual Bayesian probit model adapting both the algorithms presented in Anceschi et al. (2024). Here, we present results obtained with Algorithm 2, which is generally more efficient when p is of the same order as n, as in this case. All the details about the derivations and comparisons of the implementations are reported in the Supplementary Material.

4. RESULTS

We assess the performance of the proposed method across different dimensions and varying upper integration limits. Without loss of generality, we focus on the case where Σ is a correlation matrix rather than a covariance matrix. The estimation methods introduced in Section 3 were implemented in `C++` and integrated into `R` via the `Rcpp` and `RcppArmadillo` package (Eddelbuettel & François, 2011; Eddelbuettel & Sanderson, 2014). The simulations were run on an Intel Core i7-14700K workstation with 32 GB of RAM.

We examine matrix dimensions $m \in \{16, 64, 128, 256, 512, 1024\}$, with a fixed lower integration limit of $-\infty \mathbb{1}_m$, and an upper limit of $u = c\mathbb{1}_m$, where c spans 20 equidistant points in $[-2, 2]$ and $\mathbb{1}_m$ denotes the m-dimensional column vector of ones. We test our proposed methodology using three types of correlation matrices: (i) random correlation matrices generated following Davies & Higham (2000), (ii) constant correlation matrices with off-diagonal entries $\rho \in \{0, 0.25, 0.50, 0.75\}$, and (iii) random correlation matrices obtained by standardization of $A^\top A$, with A $m \times m$ matrix with independent standard Gaussian entries. Results for the expectation propagation approximation of $\Phi_m(u; \Sigma)$ presented in Section 3 are shown using the Cholesky factorization of the matrix $\tilde{\Sigma}$ introduced in Section 2, with $\epsilon = 0.01$. Using the eigendecomposition factorization of $\tilde{\Sigma}$ or different implementations of expectation propagation does not alter the estimates, up to numerical precision. See the Supplementary Material for details.

In each scenario, we benchmark the proposed method to three state-of-the-art algorithms: Botev's method (Botev, 2017), as implemented in the `TruncatedNormal` package, Ridgway's method (Ridgway, 2016) (authors' implementation), and — depending on the case considered — Genz's (Genz, 1992) or tile-low-rank (Cao et al., 2021) algorithms. Specifically, the tile-low-rank

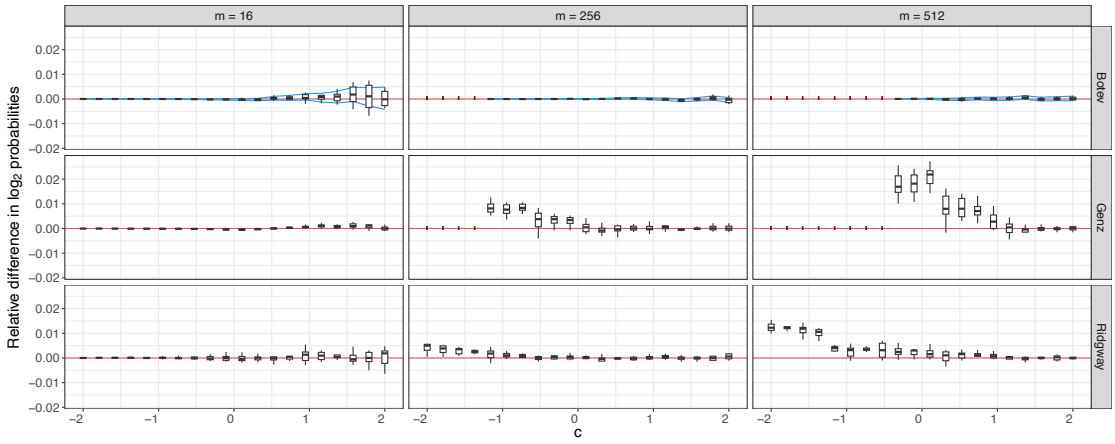

Fig. 1: Boxplots of the relative differences in the estimates of $\log_2 \Phi_m(u;\Sigma)$ obtained by various methods are shown as a function of the upper integration limit $u = c\mathbb{1}_m$, for the case where Σ is a randomly generated dense correlation matrix following Davies & Higham (2000) (case (i)). The three sampling-based methods by Botev (2017), Genz (1992), and Ridgway (2016) are benchmarked against the proposed expectation propagation approximation using the Cholesky decomposition of the matrix $\tilde{\Sigma} = \Sigma - \epsilon\lambda_m I_m$. Each boxplot summarizes the results from ten independent runs. Numerical estimates equal to $-\infty$ are marked with a vertical red tick. The blue lines indicate the mean relative error estimates obtained with the method by Botev (2017).

approximation requires the matrix Σ to have a low-rank structure, and thus finds its ideal setting in case (ii). On the other hand, when such a structure is missing, like in cases (i) and (iii), such an approximation may not be feasible and thus Genz's method is used instead. Both methods are implemented with extreme efficiency via the `pmvn()` function in the `R` package `tlrmvnmvt` (Cao et al., 2022). Each algorithm is run ten times, using 10^4 samples for sampling methods. Here, we present a subset of the results, which are completely reported in Section S.3 of the Supplementary Material. Specifically, we focus on case (i) and case (ii) with $\rho = 0.50$, when $m \in \{16, 256, 512\}$. These are a class of limited, although illustrative, examples adapted from Cao et al. (2022).

Figure 1 shows the results for case (i). For varying upper integration limit u, we plot the relative differences between the $\log_2$ estimates of $\Phi_m(u;\Sigma)$ obtained by the competitors and the ones obtained with the proposed method, which in this scenario is used as benchmark due to the possible underflow of the alternatives. The estimated values are consistent across algorithms for lower dimensions (m = 16-128, see also Figure S.3), with estimates from expectation propagation being almost indistinguishable from Botev's ones—with the expectation propagation values falling within the average estimated Botev's error bands, represented by the blue lines—and showing negligible differences with the other methods. However, a crucial difference is observed at higher dimensions, where both Botev's and Genz's methods exhibit numerical underflow for tail probabilities, resulting in estimates equal to $-\infty$. In such a case, the only viable options are Ridgway's method and expectation propagation, with the latter requiring only a fraction of the computational time (see Figure S.12), thus representing the preferable option. An analogous trend is observed for case (iii), where underflow occurs even for $m \geq 128$ (see Figure S.8). Interestingly, in such cases, when no underflow occurs for Botev's method, expectation propagation shows a lower bias than the other asymptotically-exact sampling approaches.

Figure 2 displays results when Σ is a correlation matrix with fixed structure (case (ii)) with ρ = 0.5. Given the specific structure of Σ, the ground truth can be easily computed numerically, as it

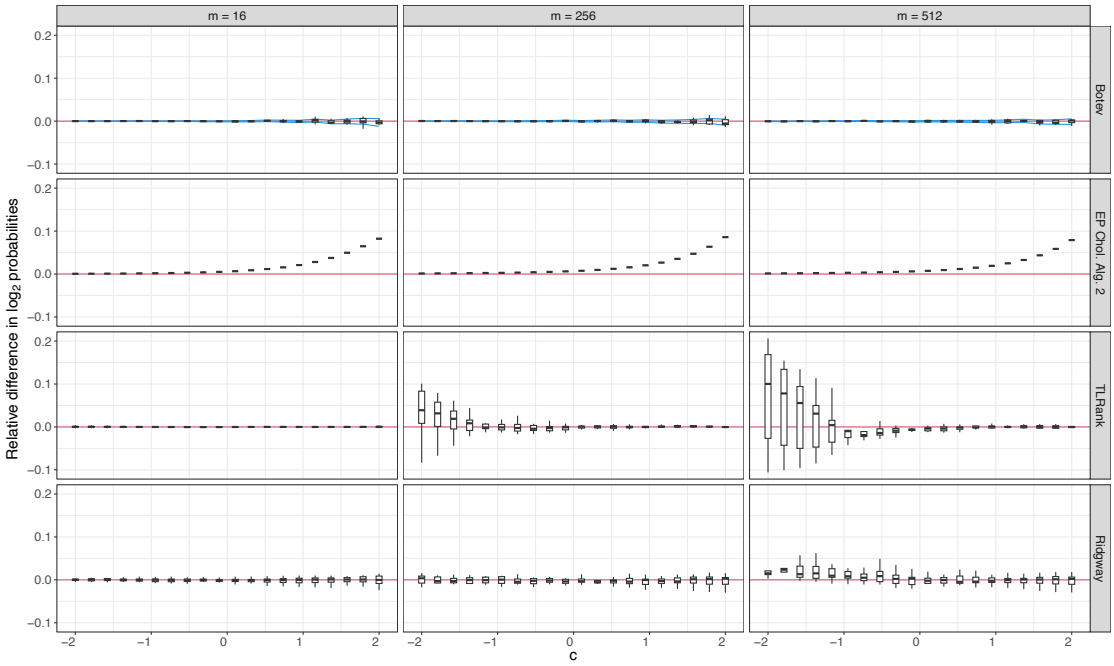

Fig. 2: Boxplots of the relative differences in the estimates of $\log_2 \Phi_m(u;\Sigma)$ obtained by various methods are shown as a function of the upper integration limit $u = c\mathbb{1}_m$ when Σ is a fixed-structure correlation matrix (case (ii)) with off-diagonal elements identically equal to $\rho = 0.50$. All four methods considered are benchmarked with the ground truth, computed numerically.

boils down to a univariate integral (see Supplementary Material). Interestingly, the tail probability estimates obtained with the proposed method align with Botev's, achieving comparable accuracy at a reduced computational cost (see Figure S.13). At the same time, the tile-low-rank algorithm (Cao et al., 2021) shows more instability in the computation of tail probabilities, empirically when the $\log_2$ probability goes below -20. On the other hand, the proposed method shows some bias in the estimation of probabilities for larger values of u, in general when the $\log_2$ probability goes above -2.5 (see Figure S.10 for the values of the $\log_2$ probabilities). This behaviour shows that expectation propagation can be the preferable option for tail probabilities, while some corrections might be useful for larger probabilities. For instance, one could resort to importance sampling using the approximate posterior as proposal or consider non-symmetric approximations (Pozza et al., 2024). Additional results for different ρ values are provided in Figures S.4, S.5, S.6, and S.7. Notably, as ρ increases, tile-low-rank gains precision and efficiency, while the accuracy of the proposed method diminishes for large values of m and u, although maintaining great precision in the computation of tail probabilities, which are usually the most challenging to compute.

The ratios between the running times of the proposed expectation propagation approach and each sampling algorithm, for the three considered cases, are reported in Figures S.12, S.13, and S.14. The proposed method is the only one computing $\Phi_m(u;\Sigma)$ in polynomial time $O(m^3)$ without sampling. Genz's method would instead require a pre-computation cost of $O(m^3)$ and $O(m^2)$ for each sample (Cao et al., 2021), with Botev's method having similar cost but higher accuracy, while Ridgway's approach would require an increased $O(m^4)$ cost. The tile-low-rank approach, when feasible thanks to the structure of Σ and implemented efficiently, has a reduced pre-processing cost of $O(m^{5/2})$, thanks to the use of a tile-low-rank Cholesky factorization, combined with a cost per sample of order $O(m^{3/2})$. In the experiments, the proposed expectation

propagation implementation is faster than the competitors for moderate dimensions, with performance comparable to `tlrmvnmvt` for $m = 256$. The latter exhibits lower running times for larger matrices ($m = 512$–1024). Yet, in high dimensions, the `tlrmvnmvt` and the `TruncatedNormal` packages may lead to underflows in the probability estimates, while Ridgway's approach may be computationally impractical, leaving expectation propagation the only viable option among the considered alternatives, especially when tail probabilities are considered. Additional studies on the trade-off between computational costs and accuracy are reported in the Supplementary Material. We also provide a link to a GitHub repository for full reproducibility.

Supplementary Material

Supplementary Material available at Biometrika online includes proofs of Proposition 1, the pseudo-code for the algorithms, and additional simulation results.

Acknowledgments

The authors are grateful to the Associate Editor and the Anonymous Reviewers for the insightful comments which considerably improved the quality of the paper.

References

Akbudak, K., Ltaief, H., Mikhalev, A. & Keyes, D. (2017). Tile low rank cholesky factorization for climate/weather modeling applications on manycore architectures. In *International Conference on High Performance Computing*. Springer.

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

Anceschi, N., Fasano, A., Durante, D. & Zanella, G. (2023a). Bayesian conjugacy in probit, tobit, multinomial probit and extensions: A review and new results. *Journal of the American Statistical Association* **118**, 1451–1469.

Anceschi, N., Fasano, A., Franzolini, B. & Rebaudo, G. (2024). Scalable expectation propagation for generalized linear models. *arXiv preprint arXiv:2407.02128* .

Anceschi, N., Fasano, A. & Rebaudo, G. (2023b). Expectation propagation for the smoothing distribution in dynamic probit. In *Bayesian Statistics, New Generations New Approaches (BAYSM 2022)*. Springer Cham.

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

Azzalini, A. (2014). *The skew-normal and related families*, vol. 3. Cambridge University Press.

Azzimonti, D. & Ginsbourger, D. (2018). Estimating orthant probabilities of high-dimensional gaussian vectors with an application to set estimation. *Journal of Computational and Graphical Statistics* **27**, 255–267.

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

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

Braunstein, A., Muntoni, A. P. & Pagnani, A. (2017). An analytic approximation of the feasible space of metabolic networks. *Nature communications* **8**, 14915.

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

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

Cao, J., Keyes, D. E., Genton, M. G. & Turkiyyah, G. M. (2022). Tlrmvnmvt: Computing high-dimensional multivariate normal and student-t probabilities with low-rank methods in r. *Journal of Statistical Software* **101**, 1–25.

Chopin, N. & Papaspiliopoulos, O. (2020). *An introduction to sequential Monte Carlo*. Springer Cham.

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

Connors, R. D., Hess, S. & Daly, A. (2014). Analytic approximations for computing probit choice probabilities. *Transportmetrica A: Transport Science* **10**, 119–139.

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

CUNNINGHAM, J. P., HENNIG, P. & LACOSTE-JULIEN, S. (2011). Gaussian probabilities and expectation propagation. *arXiv preprint arXiv:1111.6832* .

DAVIES, P. I. & HIGHAM, N. J. (2000). Numerically stable generation of correlation matrices and their factors. *Bit Numerical Mathematics* **40**, 640–651.

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

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

DING, P., IMBENS, G., QU, Z. & YE, Y. (2024). Computationally efficient estimation of large probit models. *arXiv preprint arXiv:2407.09371* .

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

EDDELBUETTEL, D. & FRANÇOIS, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software* **40**, 1–18.

EDDELBUETTEL, D. & SANDERSON, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis* **71**, 1054–1063.

FASANO, A., ANCESCHI, N., FRANZOLINI, B. & REBAUDO, G. (2023). Efficient expectation propagation for posterior approximation in high-dimensional probit models. In *Book of the Short Papers - SIS 2023*. Pearson.

FASANO, A. & DURANTE, D. (2022). A class of conjugate priors for multinomial probit models which includes the multivariate normal one. *Journal of Machine Learning Research* **23**, 1–16.

GENTON, M. G. (2004). *Skew-elliptical distributions and their applications: a journey beyond normality*. CRC Press.

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

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

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

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

MARY, T. (2017). *Block Low-Rank multifrontal solvers: complexity, performance, and scalability*. Ph.D. thesis, Université Paul Sabatier-Toulouse III.

MINKA, T. P. (2001a). Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, USA: Morgan Kaufmann Publishers Inc.

MINKA, T. P. (2001b). *A Family of Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, Massachusetts Institute of Technology.

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

OCHI, Y. E. & PRENTICE, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* **71**, 531–543.

PETERSEN, K. B., PEDERSEN, M. S. et al. (2008). The matrix cookbook. *Technical University of Denmark* **7**, 510.

POZZA, F., DURANTE, D. & SZABO, B. (2024). Skew-symmetric approximations of posterior distributions. *arXiv preprint arXiv:2409.14167* .

RASMUSSEN, C. E. & WILLIAMS, C. K. (2006). *Gaussian processes for machine learning*, vol. 2. MIT press Cambridge, MA.

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

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

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

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

WEISBECKER, C. (2013). *Improving multifrontal solvers by means of algebraic block low-rank representations*. Ph.D. thesis, Institut National Polytechnique de Toulouse-INPT.

ZHANG, C., ARRIDGE, S. & JIN, B. (2019). Expectation propagation for Poisson data. *Inverse Problems* **35**, 1–27.

ZHOU, J., ORMEROD, J. T. & GRAZIAN, C. (2023). Fast expectation propagation for heteroscedastic, lasso-penalized, and quantile regression. *Journal of Machine Learning Research* **24**, 1–39.

[*Received on* 2 *January* 2017. *Editorial decision on* 1 *August* 2023]

Biometrika (2025), **xx**, x, *pp*. 1–18
Printed in Great Britain

Supplementary Material for "Multivariate Gaussian cumulative distribution functions as the marginal likelihood of their dual Bayesian probit models"

By AUGUSTO FASANO
Dept. of Economics, Social Studies, Applied Mathematics and Statistics, University of Torino, Corso Unione Sovietica 218 bis, 10134 Torino, Italy, and Collegio Carlo Alberto
augusto.fasano@unito.it

and FRANCESCO DENTI
Dept. of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy
francesco.denti@unipd.it

Summary

The Supplementary Material includes proofs of the theoretical results, a self-contained presentation of expectation propagation for the probit model and additional simulations for the performance of the methods developed in the article "Multivariate Gaussian cumulative distribution functions as the marginal likelihood of their dual Bayesian probit models".

S.1. Proof of Proposition 1

Adapting Corollary 3 in Durante (2019), the form of the marginal likelihood for a Bayesian probit model (2) equals

$$p(y) = \Phi_n\left(DX\xi;\ \ I_n + DX\Omega(DX)^\top\right), \tag{S.1}$$

where $D = \text{diag}(2y_1 - 1, \ldots, 2y_n - 1)$.

This can also be obtained leveraging on the latent variable interpretation model (2) (Albert & Chib, 1993):

$$y_i = \mathbb{1}(z_i \geq 0), \quad z_i = x_i^\top\beta + \epsilon_i, \quad \beta \sim \mathcal{N}_p\left(\xi, \Omega\right), \tag{S.2}$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$ independent of β. Thus, *a priori*, marginally $z = (z_1, \ldots, z_n)^\top \sim \mathcal{N}_n(X\xi, I_n + X\Omega X^\top)$. For a given sample $y = (y_1, \ldots, y_n)$, call $\mathcal{A}_y = \mathcal{A}_{y_1} \times \ldots \times \mathcal{A}_{y_n}$ the cartesian product of sets $\mathcal{A}_{y_i} = \{\zeta \in \mathbb{R}\colon\ (2y_i - 1)\zeta > 0\}$, so that

$$\mathcal{A}_{y_i} = \begin{cases} (-\infty, 0) & \text{if } y_i = 0, \\ (0, +\infty) & \text{if } y_i = 1. \end{cases}$$

Thus, taking D as above and $\tilde{z} \sim \mathcal{N}_n(X\xi, I_n + X\Omega X^\top)$, it holds that the marginal likelihood for model (2) is, coherently with Durante (2019),

$$\begin{aligned} p(y) &= \mathbb{P}\left[\tilde{z} \in \mathcal{A}_y\right] = \mathbb{P}\left[D\tilde{z} \geq 0\right] = \mathbb{P}\left[D\tilde{z} - DX\xi \geq -DX\xi\right] \\ &= \mathbb{P}\left[-D\tilde{z} + DX\xi \leq DX\xi\right] = \Phi_n\left(DX\xi;\ \ I_n + DX\Omega(DX)^\top\right), \end{aligned} \tag{S.3}$$

since $-D\tilde{z} + DX\xi \sim \mathcal{N}_n(0, I_n + DX\Omega(DX)^\top)$.

Now, suppose we are interested in computing $\Phi_m(u;\ \Sigma)$, for a generic positive-definite $m \times m$ covariance matrix Σ and an arbitrary point $u \in \mathbb{R}^m$. First, let us define $\tilde{\Sigma} = \Sigma - \epsilon\lambda_m I_m$ where $\epsilon \in (0,1)$ and λ_m is the smallest eigenvalue of Σ. By construction, the matrix $\tilde{\Sigma}$ remains positive definite: the eigenvalues are obtained by subtracting $\epsilon\lambda_m$ from the eigenvalues of Σ. Being a symmetric positive-definite $m \times m$ matrix, $\tilde{\Sigma}$ can be decomposed as $\tilde{\Sigma} = P\tilde{\Lambda}P^\top$, with P and $\tilde{\Lambda}$ $m \times m$ matrices as in Proposition 1, in different possible ways (Petersen et al., 2008). For any such factorization of $\tilde{\Sigma}$, we have $\Sigma = \epsilon\lambda_m I_m + P\tilde{\Lambda}P^\top$, leading to

$$\begin{aligned}\Phi_m(u;\ \Sigma) &= \Phi_m\left(u;\ \epsilon\lambda_m I_m + P\tilde{\Lambda}P^\top\right) = \Phi_m\left(\sqrt{\frac{1}{\epsilon\lambda_m}}u;\ I_m + P\left[\frac{1}{\epsilon\lambda_m}\tilde{\Lambda}\right]P^\top\right)\\ &= \Phi_m\left(PP^{-1}\sqrt{\frac{1}{\epsilon\lambda_m}}u;\ I_m + P\left[\frac{1}{\epsilon\lambda_m}\tilde{\Lambda}\right]P^\top\right).\end{aligned} \tag{S.4}$$

Comparing (S.4) with (S.1), we can see that the generic $\Phi_m(u;\ \Sigma)$ coincides with the marginal likelihood of a dual Bayesian probit model where $n = p = m$ and relevant quantities given by

$$DX = P, \qquad \Omega = \frac{1}{\epsilon\lambda_m}\tilde{\Lambda}, \qquad \xi = P^{-1}\sqrt{\frac{1}{\epsilon\lambda_m}}u.$$

By left multiplying the first equality by D, we get that the working matrix of covariates in the dual probit must equal $X = DP$ since $D^2 = I_m$. We can also notice that in the dual probit model D and X are not uniquely identified, as they only need to satisfy the condition $DX = P$. This translates into the freedom of choosing D (or, equivalently, the working observations y) in the dual probit model, and then taking $X = DP$. For simplicity, we took $D = I_m$. For this specification, we obtain the results reported in Proposition 1.

S.2. Expectation propagation for the probit model

Here, we report the details about the expectation propagation algorithm and its specification to Bayesian probit models of the form (2). These derivations can be obtained by specializing the results in Anceschi et al. (2024) to the probit model, considering a generic prior mean ξ and covariance matrix Ω for β. See also Seeger et al. (2007); Seeger (2008); Vehtari et al. (2020); Zhou et al. (2023) for related topics. This self-contained treatment of the topic is reported to clarify the details of the employed algorithm to compute (1) also to readers unfamiliar with the method. The key idea of expectation propagation (Minka, 2001a,b) is to approximate the posterior distribution $p(\beta \mid y) \propto p(\beta)\prod_{i=1}^n p(y_i \mid \beta)$ of a p-dimensional parameter β with a density $q_{\text{EP}}(\beta)$ that has the same factorization $q_{\text{EP}}(\beta) \propto q_0(\beta)\prod_{i=1}^n q_i(\beta)$. To have a treatable global approximation $q_{\text{EP}}(\beta)$, when β is a real-valued vector, the factors (also referred to as *sites*) $q_i(\beta)$, $i = 0, \ldots, n$, are usually taken from a Gaussian-like kernel, that is $q_i(\beta) = Z_i^{-1}\exp\{-0.5\beta^\top Q_i\beta + \beta^\top r_i\}$ for $i = 0, \ldots, n$. As a consequence of this choice the global approximation takes the form $q_{\text{EP}}(\beta) = \Psi(r_{\text{EP}}, Q_{\text{EP}})^{-1}\exp\{-0.5\beta^\top Q_{\text{EP}}\beta + \beta^\top r_{\text{EP}}\}$, where $r_{\text{EP}} = \sum_{i=0}^n r_i$, $Q_{\text{EP}} = \sum_{i=0}^n Q_i$ and $\Psi(r,Q) = \int \exp\{-0.5\beta^\top Q\beta + \beta^\top r\}\, d\beta$, meaning $\log\Psi(r,Q) = 0.5\{r^\top Q^{-1}r + p\log(2\pi) - \log|Q|\}$. Thus, the density of the global approximation is a p-variate Gaussian density with mean $\xi_{\text{EP}} = Q_{\text{EP}}^{-1}r_{\text{EP}}$ and variance-covariance matrix $\Omega_{\text{EP}} = Q_{\text{EP}}^{-1}$. Now, let us denote with $\phi_p(a;B)$ the probability density function of a p-variate Gaussian with mean 0 and variance-covariance matrix B, evaluated in a. Also, we will write $\phi(a;b) \equiv \phi_1(a;b)$, $\phi(a) \equiv \phi(a;1)$, and equivalently $\Phi(a) \equiv \Phi_1(a;1)$. When a multivariate

Gaussian prior $p(\beta) = \phi_p(\beta - \xi; \Omega)$ is assumed, as for model (2), one can set $q_0(\beta) = p(\beta)$, which means taking $Q_0 = \Omega^{-1}$, $r_0 = \Omega^{-1}\xi$, and $Z_0 = \Psi(r_0, Q_0)$.

The parameters r_i, Q_i, and Z_i for the other sites, $i = 1, \ldots, n$, are instead iteratively refined via the expectation propagation recursion. In the estimation scheme, such sites play the role of approximations for the likelihood terms $p(y_i \mid \beta)$, $i = 1, \ldots, n$, and thus are not density functions in β. For this reason, also the term Z_i enters among the parameters to be optimized. See also Anceschi et al. (2024) for additional details about expectation propagation for Bayesian generalized linear models. At each iteration, until convergence is met, each site $i = 1, \ldots, n$ is updated according to the following scheme. Keeping all the other sites $q_j(\beta)$, $j \neq i$, fixed, one first computes the *cavity density* $q_{-i}(\beta) \propto q_{\text{EP}}(\beta)/q_i(\beta)$, obtaining

$$q_{-i}(\beta) = \phi_p(\beta - Q_{-i}^{-1} r_{-i}; Q_{-i}^{-1}),$$

where $r_{-i} = \sum_{j \neq i} r_j$ and $Q_{-i} = \sum_{j \neq i} Q_j$. Then, one computes the *hybrid density* $h_i(\beta) = Z_{h_i}^{-1} p(y_i \mid \beta) q_{-i}(\beta)$, with $Z_{h_i} = \int p(y_i \mid \beta) q_{-i}(\beta) d\beta$. Finally, parameters $\{r_i, Q_i, Z_i\}$ of site $q_i(\beta)$ are updated so that the resulting moments of order "zero", one, and two of the updated global approximation $q_{\text{EP}}^{\text{NEW}}(\beta) = (Z_{\text{EP}}^{\text{NEW}})^{-1} q_i^{\text{NEW}}(\beta) q_{-i}(\beta)$, with $Z_{\text{EP}}^{\text{NEW}} = \int q_{-i}(\beta) q_i^{\text{NEW}}(\beta) d\beta = \Psi(r_{\text{EP}}^{\text{NEW}}, Q_{\text{EP}}^{\text{NEW}}) / \{\Psi(r_{-i}, Q_{-i}) \cdot Z_i^{\text{NEW}}\}$, match the ones of the hybrid distribution, that is

$$\begin{cases} Z_{\text{EP}}^{\text{NEW}} = Z_{h_i} \\ \left(Q_{-i} + Q_i^{\text{NEW}}\right)^{-1} \left(r_{-i} + r_i^{\text{NEW}}\right) = \mu_{h_i} \\ \left(Q_{-i} + Q_i^{\text{NEW}}\right)^{-1} = \Sigma_{h_i}, \end{cases}$$

where $\mu_{h_i} = \mathbb{E}_{h_i(\beta)}[\beta]$ and $\Sigma_{h_i} = \text{var}_{h_i(\beta)}[\beta]$. This leads to

$$\begin{cases} \log Z_i^{\text{NEW}} = \log \Psi(r_{\text{EP}}^{\text{NEW}}, Q_{\text{EP}}^{\text{NEW}}) - \log \Psi(r_{-i}, Q_{-i}) - \log Z_{h_i} \\ r_i^{\text{NEW}} = \left(Q_{-i} + Q_i^{\text{NEW}}\right) \mu_{h_i} - r_{-i} \\ Q_i^{\text{NEW}} = \Sigma_{h_i}^{-1} - Q_{-i}. \end{cases}$$

After the algorithm has converged, the log-marginal likelihood $\log p(y)$ can be approximated as

$$\log m_{\text{EP}}(y) = \log \Psi(r_{\text{EP}}, Q_{\text{EP}}) - \log \Psi(r_0, Q_0) - \sum_{i=1}^{n} \log Z_i. \tag{S.5}$$

The applicability of this approach is strictly related to the computation of the normalizing constant Z_{h_i} and the hybrid moments μ_{h_i} and Σ_{h_i}. This depends on the tractability of the hybrid distribution $h_i(\beta)$. In the case of the probit model, the hybrid density is the density of a multivariate extended skew-normal distribution $\text{SN}_p(\xi_i, \Omega_i, \alpha_i, \tau_i)$ (see Azzalini (2014)), with

$$\xi_i = Q_{-i}^{-1} r_{-i}, \quad \Omega_i = Q_{-i}^{-1}, \quad \alpha_i = (2y_i - 1)\omega_i x_i, \quad \tau_i = (2y_i - 1)(1 + x_i^\top \Omega_i x_i)^{-1/2} x_i^\top \xi_i,$$

where $\omega_i = [\text{diag}(\Omega_i)]^{1/2}$. Hence, all the needed quantities are available in closed form, having $Z_{h_i} = \Phi(\tau_i)$, $\mu_{h_i} = \xi_i + \zeta_1(\tau_i) s_i \Omega_i x_i$, and $\Sigma_{h_i} = \Omega_i + \zeta_2(\tau_i) s_i^2 (\Omega_i x_i)(\Omega_i x_i)^\top$, with $s_i = (2y_i - 1)(1 + x_i^\top \Omega_i x_i)^{-1/2}$, $\zeta_1(x) = \phi(x)/\Phi(x)$ and $\zeta_2(x) = -\zeta_1(x)^2 - x\zeta_1(x)$. Moreover, it can be shown that both r_i and Q_i are parameterized by scalar quantities, as $r_i = m_i x_i$ and $Q_i = k_i x_i x_i^\top$, for some real quantities m_i and k_i, $i = 1, \ldots, n$ (Anceschi et al., 2024).

Thus, for each site, one has to update only the three scalar quantities $\{m_i, k_i, \log Z_i\}$ (see also Seeger et al. (2007); Seeger (2008)), for which the following holds

$$\begin{cases} k_i^{\text{NEW}} = -\zeta_2(\tau_i)/\left(1 + x_i^\top \Omega_i x_i + \zeta_2(\tau_i) x_i^\top \Omega_i x_i\right) , \\ m_i^{\text{NEW}} = \zeta_1(\tau_i) s_i + k_i^{\text{NEW}} x_i^\top \Omega_i r_{-i} + k_i^{\text{NEW}} \zeta_1(\tau_i) s_i x_i^\top \Omega_i x_i , \\ \log Z_i^{\text{NEW}} = \dfrac{1}{2}\left[\dfrac{2 m_i^{\text{NEW}} r_{-i}^\top \Omega_i x_i + (m_i^{\text{NEW}})^2 x_i^\top \Omega_i x_i - k_i^{\text{NEW}} (r_{-i}^\top \Omega_i x_i)^2}{1 + k_i^{\text{NEW}} x_i^\top \Omega_i x_i} - \log(1 + k_i^{\text{NEW}} x_i^\top \Omega_i x_i) \right] \\ \qquad\qquad - \log \Phi(\tau_i). \end{cases}$$

The matrix $\Omega_i = \left(Q_{\text{EP}} - k_i x_i x_i^\top\right)^{-1}$ needed for the updates can be computed exploiting Woodbury's identity as

$$\Omega_i = \Omega_{\text{EP}} + \frac{k_i}{1 - k_i x_i^\top \Omega_{\text{EP}} x_i} \left(\Omega_{\text{EP}} x_i\right) \left(\Omega_{\text{EP}} x_i\right)^\top ,$$

where the expectation propagation global covariance matrix Ω_{EP} is updated at each iteration exploiting the moment matching conditions $\Omega_{\text{EP}}^{\text{NEW}} = (Q_{-i} + Q_i^{\text{NEW}})^{-1} = \Sigma_{h_i}$. After convergence is reached, the approximation of the marginal likelihood is obtained via (S.5) as

$$\log m_{\text{EP}}(y) = \frac{1}{2}\left[r_{\text{EP}}^\top \xi_{\text{EP}} - \log |Q_{\text{EP}}| - r_0^\top \xi - \log |\Omega| \right] - \sum_{i=1}^{n} \log Z_i .$$

When the prior covariance Ω is diagonal, as in the considered dual probit models, $\log |\Omega|$ simply equals the sum of the logarithms of the diagonal elements. In addition, since usually k_i and m_i, $i = 1, \ldots, n$, are initialized to zero, so that the initial expectation propagation approximation matches the prior distribution, one can initialize $\log |Q_{\text{EP}}|$ to $-\log |\Omega|$ and then simply update $\log |Q_{\text{EP}}|$ at each iteration exploiting the equality $\log |Q_{\text{EP}}^{\text{NEW}}| = \log \left|Q_{\text{EP}}\right| + \log \left[1 + (k_i^{\text{NEW}} - k_i) x_i^\top \Omega_{\text{EP}} x_i\right]$, where k_i and k_i^{NEW} represent values before and after site i is updated, respectively.

S.2.1. *Two possible implementations for the Bayesian probit model*

Performing the updates reported above, one would obtain an implementation for posterior inference in a probit model (2) having cost $O(p^2 n)$ per iteration, corresponding to the specification of Algorithm 1 in Anceschi et al. (2024) to the case at hand. This is reported in Algorithm 1 below for completeness.

Algorithm 1. Expectation propagation for probit model (2) - $O(p^2n)$ cost per iteration

Initialization:

$\Omega_{\text{EP}} = \Omega$; $r_{\text{EP}} = r_0 = \Omega^{-1}\xi$; $\log|Q_{\text{EP}}| = -\log|\Omega|$;

$k_i = 0$, $m_i = 0$, and $\log Z_i = 0$ for $i = 1, \dots, n$.

Optimization:

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

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

    ..... Cavity distribution

    $\Omega_i = \Omega_{\text{EP}} + k_i/\left(1 - k_i x_i^\top \Omega_{\text{EP}} x_i\right) (\Omega_{\text{EP}} x_i)\,(\Omega_{\text{EP}} x_i)^\top$

    $r_{-i} = r_{\text{EP}} - m_i x_i$

    ..... Hybrid distribution

    $s_i = (2y_i - 1)(1 + x_i^\top \Omega_i x_i)^{-1/2}$

    $\tau_i = s_i x_i^\top \Omega_i r_{-i}$

    ..... i-th site approximation

    $k_i^{\text{NEW}} = -\zeta_2(\tau_i)/\left(1 + x_i^\top \Omega_i x_i + \zeta_2(\tau_i) x_i^\top \Omega_i x_i\right)$; $\quad \delta_{k_i} = k_i^{\text{NEW}} - k_i$; $\quad k_i = k_i^{\text{NEW}}$

    $m_i = \zeta_1(\tau_i) s_i + k_i x_i^\top \Omega_i r_{-i} + k_i \zeta_1(\tau_i) s_i x_i^\top \Omega_i x_i$

    $$\log Z_i = \frac{1}{2}\left[\frac{2m_i r_{-i}^\top \Omega_i x_i + m_i^2 x_i^\top \Omega_i x_i - k_i (r_{-i}^\top \Omega_i x_i)^2}{1 + k_i x_i^\top \Omega_i x_i} - \log(1 + k_i x_i^\top \Omega_i x_i)\right] - \log \Phi(\tau_i)$$

    ..... Global approximation

    $r_{\text{EP}} = r_{-i} + m_i x_i$

    $\Omega_{\text{EP}} = \Omega_i + \zeta_2(\tau_i) s_i^2 (\Omega_i x_i)(\Omega_i x_i)^\top$

    $\log|Q_{\text{EP}}| = \log|Q_{\text{EP}}| + \log[1 + \delta_{k_i} x_i^\top \Omega_{\text{EP}} x_i]$

Final computation of quantities of interest:

$\xi_{\text{EP}} = \Omega_{\text{EP}} r_{\text{EP}}$

$$\log m_{\text{EP}}(y) = \frac{1}{2}\left[\, r_{\text{EP}}^\top \xi_{\text{EP}} - \log|Q_{\text{EP}}| - r_0^\top \xi - \log|\Omega| \,\right] - \textstyle\sum_{i=1}^n \log Z_i$$

Output: $(\xi_{\text{EP}}, \Omega_{\text{EP}}, \log m_{\text{EP}}(y))$

An alternative implementation is also possible, with cost $O(pn^2)$ per iteration. This is faster when $p > n$ and generally more efficient when p has the same order as n. If using expectation propagation to compute a Gaussian cumulative distribution function (1), the dual probit model has $n = p = m$, so both implementations would have $O(m^3)$ cost. However, this alternative implementation is usually more efficient when p has the same order of magnitude as n and, for this reason, will be preferred. See also Figure S.2 for reference. We highlight that these are different implementations of the same method, so the difference is merely computational. Results are indeed the same up to numerical precision, see Figure S.1. Should more efficient implementations become available, the conclusions about the accuracy of the method would still be valid. In fact, the expectation propagation algorithm can be written in terms of the p-dimensional vectors $w_i = \Omega_i x_i$ and $v_i = \Omega_{\text{EP}} x_i$, $i = 1, \dots, n$, instead of working with the $p \times p$ matrices Ω_i and Ω_{EP}. It holds $w_i = d_i v_i$, with $d_i = (1 - k_i x_i^\top v_i)^{-1}$. Each time a site i is updated, the resulting global covariance matrix Ω_{EP} changes, and thus all the v_j's, $j = 1, \dots, n$ need to be updated according to the rule $v_j^{\text{NEW}} = v_j - c_i (x_i^\top v_j) v_i$, with $c_i = (k_i^{\text{NEW}} - k_i)/(1 + (k_i^{\text{NEW}} - k_i) x_i^\top v_i)$. Following Anceschi et al. (2024), one can define the $p \times n$ matrix $V = [v_1, v_2, \dots, v_n] = \Omega_{\text{EP}} X^\top$, and update it as

$$V^{\text{NEW}} = V - c_i v_i x_i^\top V. \tag{S.6}$$

Finally, after convergence, one can recover Ω_{EP} as $\Omega_{\text{EP}} = \Omega - VKX\Omega$, where $K = \text{diag}(k_1, \ldots, k_n)$. Combining the above, one obtains the expectation propagation implementation in Algorithm 2 below, which corresponds to Algorithm 2 in Anceschi et al. (2024), specified to the probit model. This is the algorithm used in our experiments to compute $\Phi_m(u; \Sigma)$ exploiting Proposition 1 and approximating the marginal likelihood of the dual probit model via expectation propagation.

Algorithm 2. Expectation propagation for probit model (2) - $O(pn^2)$ cost per iteration

Initialization:

$r_{\text{EP}} = r_0 = \Omega^{-1}\xi$; $V = [v_1, \ldots, v_n] = \Omega X^\top$; $\log|Q_{\text{EP}}| = -\log|\Omega|$;

$k_i = 0$, $m_i = 0$, and $\log Z_i = 0$ for $i = 1, \ldots, n$.

Optimization:

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

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

    Cavity distribution

    $w_i = (1 - k_i x_i^\top v_i)^{-1} v_i$

    $r_{-i} = r_{\text{EP}} - m_i x_i$

    Hybrid distribution

    $s_i = (2y_i - 1)(1 + x_i^\top w_i)^{-1/2}$

    $\tau_i = s_i w_i^\top r_{-i}$

    i-th site approximation

    $k_i^{\text{NEW}} = -\zeta_2(\tau_i)/\big(1 + x_i^\top w_i + \zeta_2(\tau_i) x_i^\top w_i\big)$; $\quad \delta_{k_i} = k_i^{\text{NEW}} - k_i$; $\quad k_i = k_i^{\text{NEW}}$

    $m_i = \zeta_1(\tau_i) s_i + k_i w_i^\top r_{-i} + k_i \zeta_1(\tau_i) s_i x_i^\top w_i$

    $$\log Z_i = \frac{1}{2}\left[\frac{2m_i r_{-i}^\top w_i + m_i^2 x_i^\top w_i - k_i (r_{-i}^\top w_i)^2}{1 + k_i x_i^\top w_i} - \log(1 + k_i x_i^\top w_i)\right] - \log\Phi(\tau_i)$$

    Global approximation

    $r_{\text{EP}} = r_{-i} + m_i x_i$

    $V = V - v_i \left[\delta_{k_i}/\big(1 + \delta_{k_i} x_i^\top v_i\big)\right] x_i^\top V$

    $\log|Q_{\text{EP}}| = \log|Q_{\text{EP}}| + \log[1 + \delta_{k_i} x_i^\top v_i]$

Final computation of quantities of interest:

$\Omega_{\text{EP}} = \Omega - VKX\Omega$, with $K = \text{diag}(k_1, \ldots, k_n)$

$\xi_{\text{EP}} = \Omega r_{\text{EP}} - VKX\Omega r_{\text{EP}}$

$$\log m_{\text{EP}}(y) = \frac{1}{2}\big[\, r_{\text{EP}}^\top \xi_{\text{EP}} - \log|Q_{\text{EP}}| - r_0^\top \xi - \log|\Omega| \,\big] - \textstyle\sum_{i=1}^n \log Z_i$$

Output: $(\xi_{\text{EP}}, \Omega_{\text{EP}}, \log m_{\text{EP}}(y))$

S.3. Additional Results

In this section, we first compare in terms of estimation accuracy and computational times the two expectation propagation Algorithms 1 and 2 presented above using both the Cholesky factorization and the eigendecomposition for $\tilde{\Sigma}$, motivating our choice for the use of Algorithm 2 combined with the Cholesky factorization. With reference to both Algorithms 1 and 2, we adopted the following convergence criterion. For each observation $i = 1, \ldots, n$, we compute relative change from the previous values of k_i, m_i, and Z_i. Finally, we stop the algorithm if the maxima of all three relative variations fall below a certain threshold, which we set conservatively to 0.0001. Then, we report additional figures to complete the exposition of the results discussed

in Section 4 of the main paper. All code used in the studies is openly available at the GitHub repository `Fradenti/DualProbitCDF`.

S.3.1. *Comparing the expectation propagation algorithms*

In Figures S.1 and S.2 we compare the results obtained by the different implementations of the expectation propagation method presented in Section S.2 across all the different simulations. The boxplots are stratified according to the correlation matrix structure (panels). As expected, the different algorithms provide almost identical results in terms of estimated probabilities, as we can appreciate from Figure S.1. That said, it is interesting to note from Figure S.2 that, albeit the different algorithms share the same theoretical scaling as function of m, the first version of our proposals (Algorithm 1) necessitates of considerably longer times to reach convergence, with a difference that exacerbates as the dimension increases.

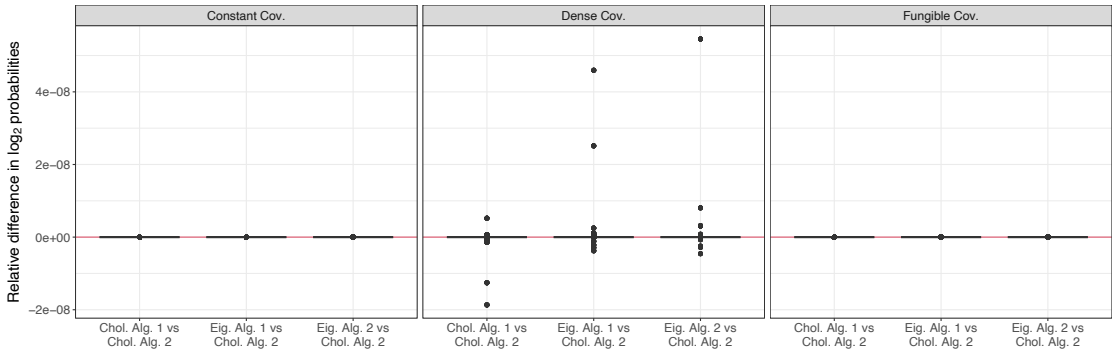

Fig. S.1: For both the expectation propagation implementations and both matrix decompositions considered, we report boxplots of the relative differences of the $\log_2$ probabilities estimates, using the version based on Cholesky decomposition and Algorithm 2 as the benchmark. These ratios are computed across all simulations, upper integration limits, and across multiple runs.

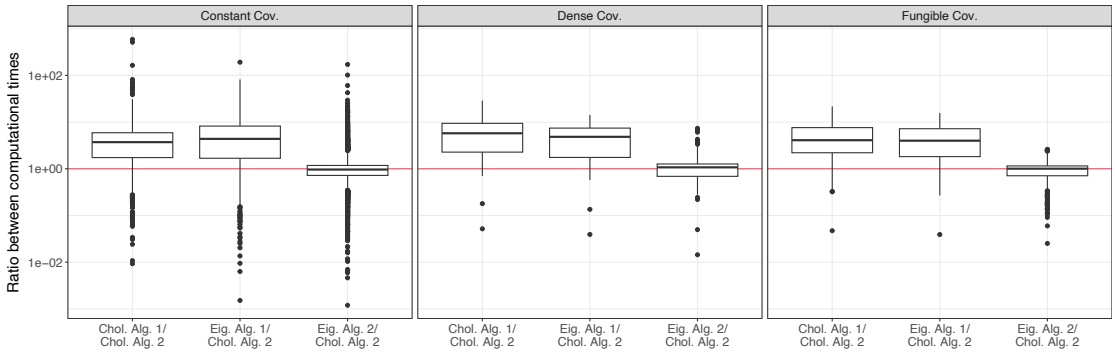

Fig. S.2: For both the expectation propagation implementations and both matrix decompositions considered, we report boxplots of the ratios of computational times, using the version based on Cholesky decomposition and Algorithm 2 as the benchmark. These ratios are computed across all simulations, upper integration limits, and multiple runs.

S.3.2. *Complete results of the simulation studies*

The following figures contain the complete results for all the dimensions $m \in \{16, 64, 128, 256, 512, 1024\}$ from our extensive simulation studies, deferred to this Supplementary Material for the sake of conciseness. In particular, Figure S.3 shows the results about case (i), where Σ is a random dense correlation matrix generated according to Davies & Higham (2000), which is partially mentioned in the main text.

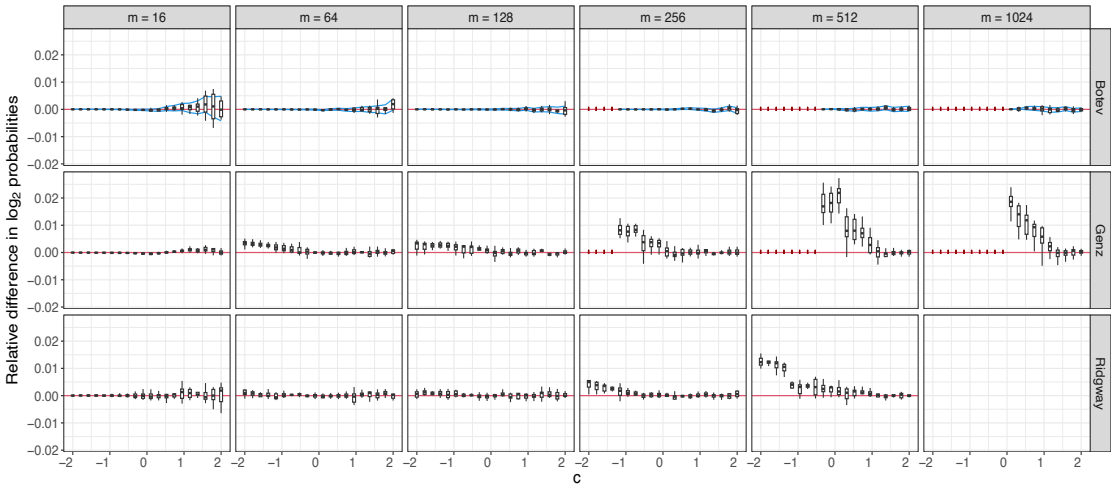

Fig. S.3: Boxplots of the relative differences in the estimates of $\log_2 \Phi_m(u; \Sigma)$ obtained by various methods are shown as a function of the upper integration limit $u = c\mathbb{1}_m$, for the case where Σ is a randomly generated dense correlation matrix following Davies & Higham (2000) (case (i)). The three sampling-based methods by Botev (2017), Genz (1992), and Ridgway (2016) are benchmarked against the proposed expectation propagation approximation using the Cholesky decomposition of the matrix $\tilde{\Sigma} = \Sigma - \epsilon\lambda_m I_m$. Each boxplot summarizes the results from ten independent runs. Numerical estimates equal to $-\infty$ are marked with a vertical red tick. The blue lines indicate the mean relative error estimates obtained with the method by Botev (2017). Results for the methods of Ridgway (2016) in dimension $m = 1024$ were not computed due to their high computational cost.

The figure shows the boxplots of relative differences in the estimates of $\log_2 \Phi_m(u; \Sigma)$ obtained with the proposed method and with competitors across ten repeated evaluations, for varying values of the upper integration limit $u = c\mathbb{1}_m$, where c spans 20 equidistant points in $[-2, 2]$ and $\mathbb{1}_m$ denotes the m-dimensional column vector of ones. Since some of the competing methods may result in uderflows for small probabilities, the proposed method is used as benchmark, due to its higher stability. Thus, for each method M, at each value of u we represent the boxplot of $(\log_2 \hat{p}_M^{(r)} - \log_2 \hat{p}_{EP})/\log_2 \hat{p}_{EP}$, $r = 1 \ldots, 10$, where $\hat{p}_M^{(r)}$ is the probability estimate obtained with the generic method M in replication r and $\hat{p}_{EP}$ is the estimate obtained with expectation propagation, for the specific value of u considered. Since, when no underflow issues are present, probability estimates given by the `TruncatedNormal` package also come with an estimate $\hat{\epsilon}$ of the relative error, we represented these estimates graphically, to assess the accuracy of Botev's method, and, consequently, of expectation propagation. More specifically, calling $\hat{p}^{(r)}$ and $\hat{\epsilon}^{(r)}$ the values of $\hat{p}$ and $\hat{\epsilon}$ obtained with Botev's method for replication $r = 1, \ldots, 10$, we get the lower and upper bounds $\hat{p}_l^{(r)} = \hat{p}^{(r)}(1 - \hat{\epsilon}^{(r)})$ and $\hat{p}_u^{(r)} = \hat{p}(1 + \hat{\epsilon}^{(r)})$ for the estimated probability. We

thus reported the mean (across the 10 replicated experiments) relative discrepancies between such quantities (transformed in $\log_2$ scale) and the benchmark, obtaining the blue bands which may be seen as a measure of the reliability of Botev's estimates. That is, for each value of u, the blue lines represent the average, across the 10 replicated experiments, of the values $(\log_2 \hat{p}_l^{(r)} - \log_2 \hat{p}_{EP})/\log_2 \hat{p}_{EP}$ and $(\log_2 \hat{p}_u^{(r)} - \log_2 \hat{p}_{EP})/\log_2 \hat{p}_{EP}$. A similar procedure is used in the other figures, with the only difference being that for case (ii) the probability of interest can easily be computed numerically and is thus used as benchmark instead of expectation propagation. Indeed, Figures S.4, S.5, S.6 and S.7 display all the $\log_2$ probability estimates under case (ii) described in the manuscript, where a correlation matrix with fixed-structure is considered. These four figures refer to the cases where the off-diagonal correlations are set to $\rho = 0$, $\rho = 0.25$, $\rho = 0.50$ (already partly mentioned in the text), and $\rho = 0.75$, respectively. In such a case, the ground truth can be easily computed numerically, as the multivariate Gaussian cumulative distribution function reduces to a univariate integral thanks to the equality (see, for instance, pages 192-193 in Tong, 1990)

$$\Phi_m(u; \Sigma) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left\{-\frac{t^2}{2}\right\} \prod_{i=1}^{m} \Phi\left(\frac{u_i + \sqrt{\rho}t}{\sqrt{1-\rho}}\right) dt.$$

Consequently, this quantity is used as benchmark in all the experiments for case (ii). Finally, Figure S.8 presents the results for case (iii), with random dense correlation matrices generated by first sampling a matrix A with independent standard Gaussian entries and then computing and standardizing $A^\top A$ to obtain a correlation matrix.

Finally, note that the algorithm presented in Ridgway (2016) includes a resampling step performed whenever the effective sample size falls below a certain threshold, followed by a move step. The threshold is set by default at 50% of the original sample size, but was reduced to 10% in our implementation to mitigate the substantial computational burden when $m = 512$. For analogous computational reasons, results for this method when $m = 1024$ were not computed.

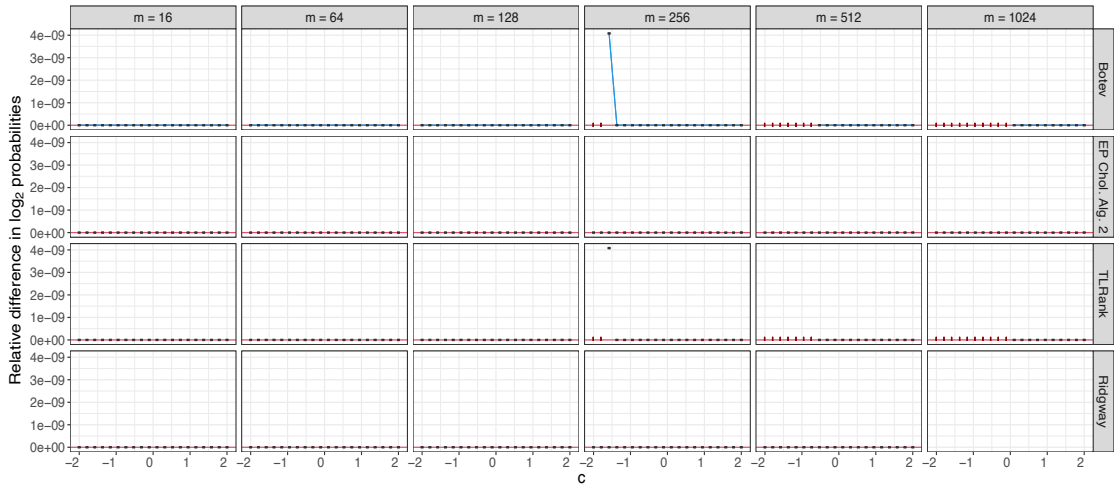

Fig. S.4: Boxplots of the relative differences in the estimates of $\log_2 \Phi_m(u;\Sigma)$ obtained by various methods are shown as a function of the upper integration limit $u = c\mathbb{1}_m$ when Σ is a diagonal correlation matrix (case (ii) with $\rho = 0$). All four methods considered are benchmarked with the ground truth, computed numerically. Each boxplot summarizes the results from ten independent runs. Numerical estimates equal to $-\infty$ are marked with a vertical red tick. The blue lines indicate the mean relative error estimates obtained with the method by Botev (2017). Results for the methods of Ridgway (2016) in dimension $m = 1024$ were not computed due to their high computational cost.

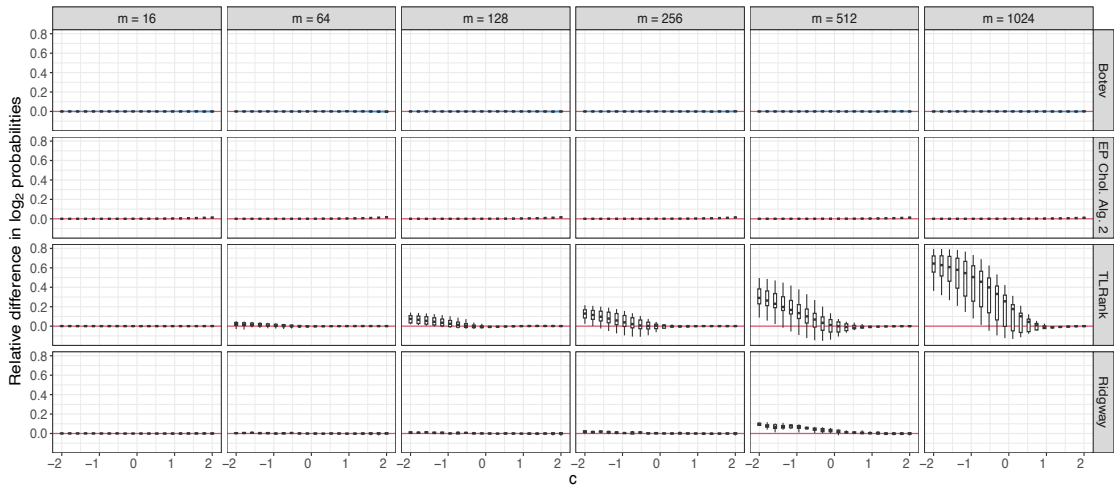

Fig. S.5: Boxplots of the relative differences in the estimates of $\log_2 \Phi_m(u;\Sigma)$ obtained by various methods are shown as a function of the upper integration limit $u = c\mathbb{1}_m$ when Σ is a fixed-structure correlation matrix (case (ii)) with off-diagonal elements identically equal to $\rho = 0.25$. All four methods considered are benchmarked with the ground truth, computed numerically. Each boxplot summarizes the results from ten independent runs. Numerical estimates equal to $-\infty$ are marked with a vertical red tick. The blue lines indicate the mean relative error estimates obtained with the method by Botev (2017). Results for the methods of Ridgway (2016) in dimension $m = 1024$ were not computed due to their high computational cost.

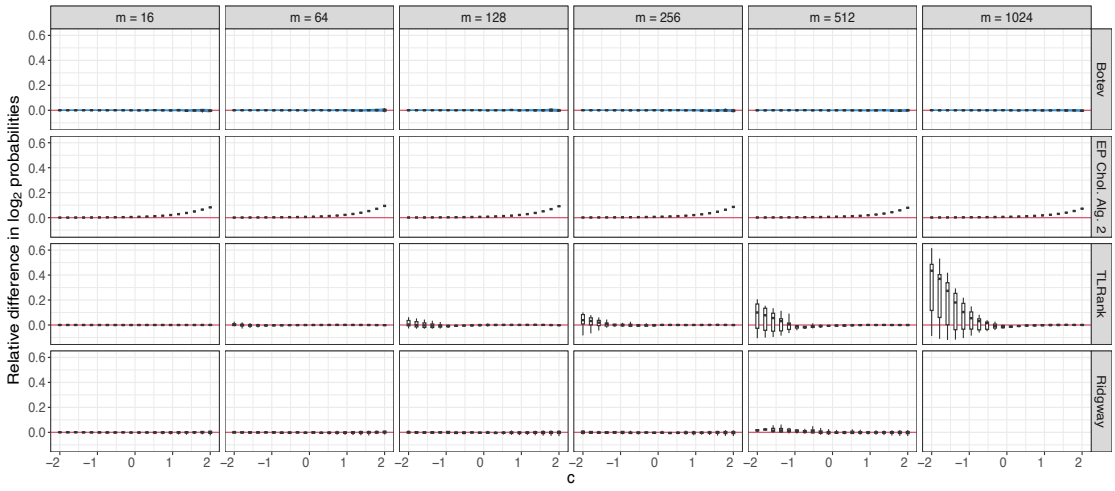

Fig. S.6: Boxplots of the relative differences in the estimates of $\log_2 \Phi_m(u;\Sigma)$ obtained by various methods are shown as a function of the upper integration limit $u = c\mathbb{1}_m$ when Σ is a fixed-structure correlation matrix (case (ii)) with off-diagonal elements identically equal to $\rho = 0.50$. All four methods considered are benchmarked with the ground truth, computed numerically. Each boxplot summarizes the results from ten independent runs. Numerical estimates equal to $-\infty$ are marked with a vertical red tick. The blue lines indicate the mean relative error estimates obtained with the method by Botev (2017). Results for the methods of Ridgway (2016) in dimension $m = 1024$ were not computed due to their high computational cost.

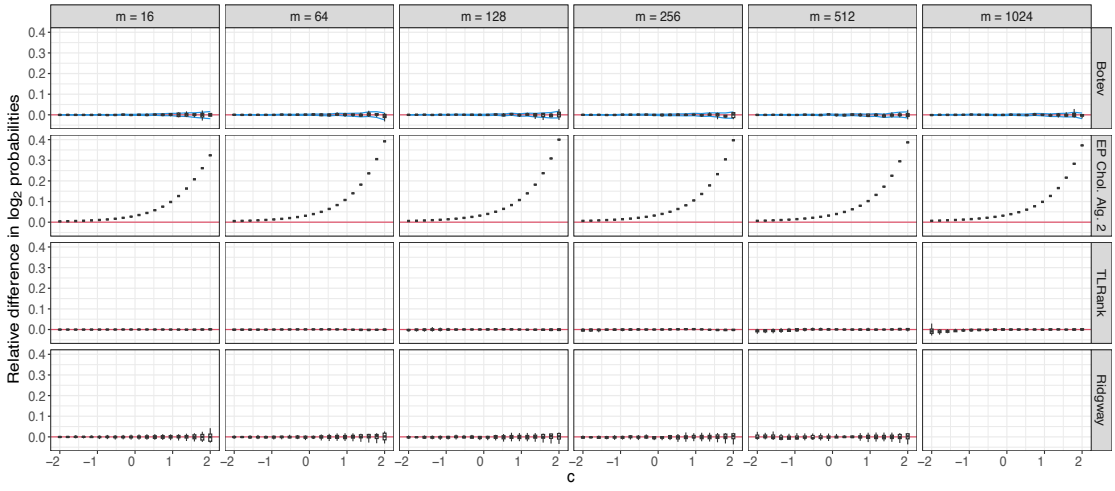

Fig. S.7: Boxplots of the relative differences in the estimates of $\log_2 \Phi_m(u;\Sigma)$ obtained by various methods are shown as a function of the upper integration limit $u = c\mathbb{1}_m$ when Σ is a fixed-structure correlation matrix (case (ii)) with off-diagonal elements identically equal to $\rho = 0.75$. All four methods considered are benchmarked with the ground truth, computed numerically. Each boxplot summarizes the results from ten independent runs. Numerical estimates equal to $-\infty$ are marked with a vertical red tick. The blue lines indicate the mean relative error estimates obtained with the method by Botev (2017). Results for the methods of Ridgway (2016) in dimension $m = 1024$ were not computed due to their high computational cost.

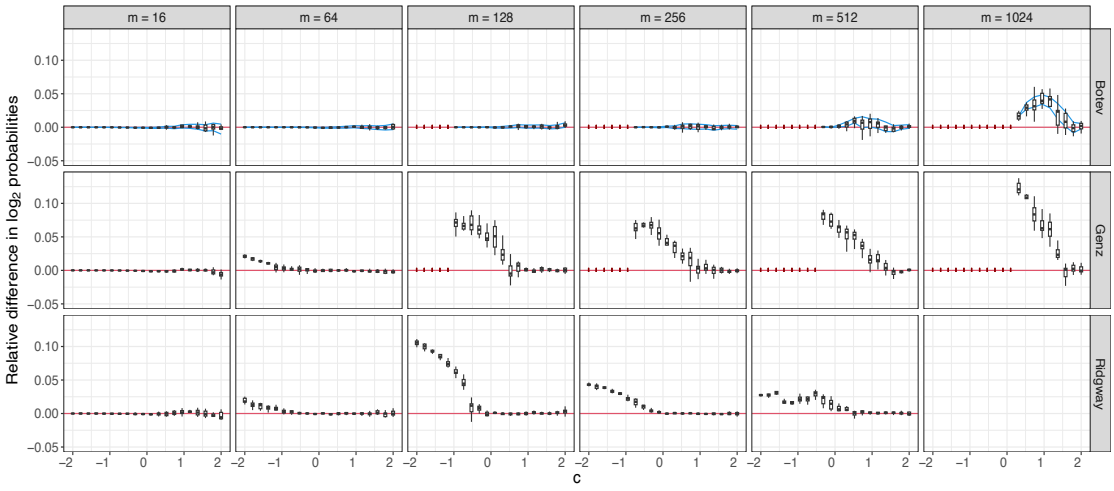

Fig. S.8: Boxplots of the relative differences in the estimates of $\log_2 \Phi_m(u;\Sigma)$ obtained by various methods as a function of the upper integration limit $u = c\mathbb{1}_m$ estimated with random dense correlation matrix (case (iii)). The three sampling-based methods by Botev (2017), Genz (1992), and Ridgway (2016) are benchmarked against the proposed expectation propagation approximation using the Cholesky decomposition of the matrix $\tilde{\Sigma} = \Sigma - \epsilon\lambda_m I_m$. Each boxplot summarizes the results from ten independent runs. Numerical estimates equal to $-\infty$ are marked with a vertical red tick. The blue lines indicate the mean relative error estimates obtained with the method by Botev (2017). Results for the methods of Ridgway (2016) in dimension $m = 1024$ were not computed due to their high computational cost.

To provide a complementary perspective, it is informative to examine how the estimated value of $\log_2 \Phi_m(u;\Sigma)$ varies as a function of the upper integration limit $u = c\mathbb{1}_m$ across the four methods considered. Figures S.9, S.10, and S.11 report results for cases (i), (ii), and (iii), respectively. Each figure stratifies the estimates by dimension $m \in \{16, 64, 128, 256, 512, 1024\}$. As previously discussed, the method by Ridgway (2016) was not run due to its high computational cost.

Each dot represents a single simulation run, while the lines connect the average estimates at each value of u. In some instances, the methods of Botev (2017) and Genz (1992) suffer from numerical underflow at smaller values of u. These cases are indicated with tick marks for estimates equal to $-\infty$, and the largest value of u at which underflow first occurs is highlighted with a vertical line, colored according to the corresponding method. Empirically, we observe that Genz's and Botev's methods may experience underflow issues when the $\log_2$-probability goes below -1000. Ridgway's approach does not experience these issues, but this comes at an increased computational cost, which makes it potentially impractical in high-dimensional settings.

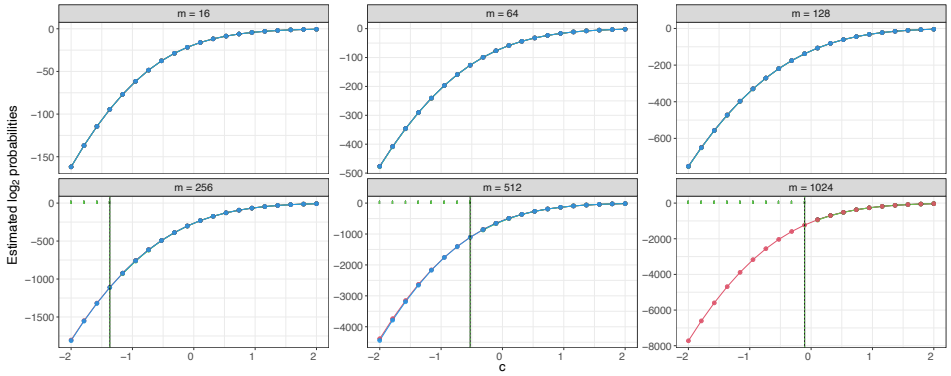

Fig. S.9: Estimates of $\log_2 \Phi_m(u;\Sigma)$ obtained by four methods as a function of the upper integration limit $u = c\mathbb{1}_m$, where Σ is a random dense correlation matrix generated following Davies & Higham (2000) (case (i)). The sampling-based methods by Botev (2017) (black), Genz (1992) (green), and Ridgway (2016) (blue; not computed for $m = 1024$) are compared with the proposed expectation propagation (EP) approximation using a Cholesky factorization (red). Each point represents a single run, and lines connect the corresponding averages. Numerical estimates equal to $-\infty$ (underflows) are indicated with ticks. Vertical lines, colored according to the method, mark the largest u values for which underflow begins to occur in the tail.

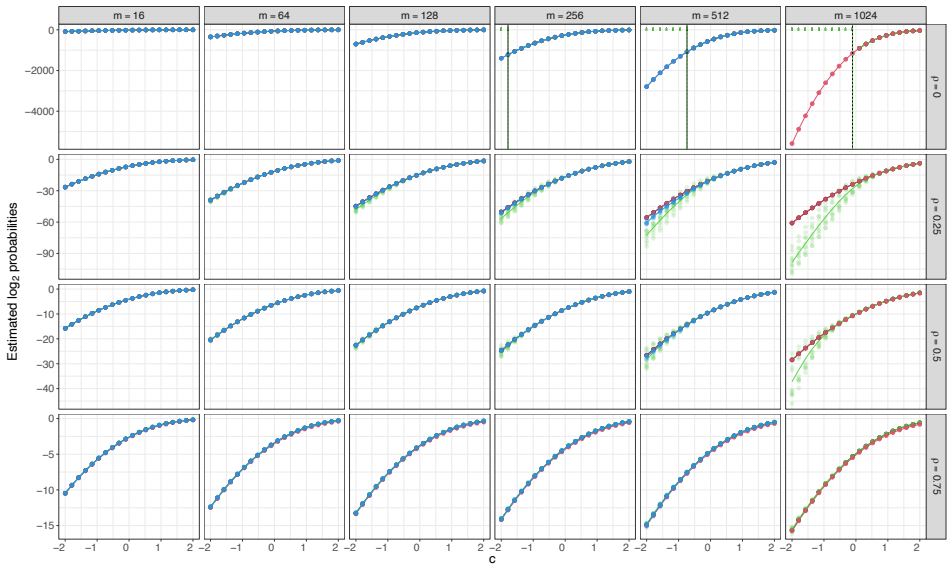

Fig. S.10: Estimates of $\log_2 \Phi_m(u; \Sigma)$ obtained by various methods as a function of the upper integration limit $u = c\mathbb{1}_m$ estimated with fixed-structure correlation matrix (case (ii)) with off-diagonal elements identically equal to ρ. The sampling-based methods by Botev (2017) (black), Cao et al. (2021) (green), and Ridgway (2016) (blue; not computed for $m = 1024$) are compared with the proposed expectation propagation (EP) approximation using a Cholesky factorization (red). Each point represents a single run, and lines connect the corresponding averages. Numerical estimates equal to $-\infty$ (underflows) are indicated with ticks. Vertical lines, colored according to the method, mark the largest u values for which underflow begins to occur in the tail.

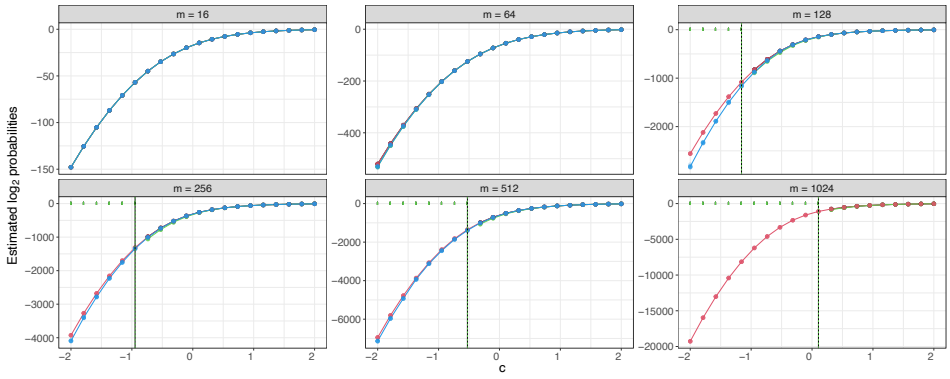

Fig. S.11: Estimates of $\log_2 \Phi_m(u; \Sigma)$, obtained by various methods as a function of the upper integration limit $u = c\mathbb{1}_m$ estimated with random dense correlation matrix generated under case (iii). The sampling-based methods by Botev (2017) (black), Genz (1992) (green), and Ridgway (2016) (blue; not computed for $m = 1024$) are compared with the proposed expectation propagation (EP) approximation using a Cholesky factorization (red). Each point represents a single run, and lines connect the corresponding averages. Numerical estimates equal to $-\infty$ (underflows) are indicated with ticks. Vertical lines, colored according to the method, mark the largest u values for which underflow begins to occur in the tail.

S.3.3. *Comparing computational costs*

Figure S.12, S.13, and S.14 present the ratio between the running times of expectation propagation and every other algorithm, stratified by dimension and correlation structure when the correlation matrices are generated under case (i), (ii), and (iii), respectively. In the experiments, the expectation propagation approximation is faster than the competitors for moderate dimensions, with performance comparable to `tlrmvnmvt` for $m = 256$. The latter exhibits lower running times for larger matrices ($m = 512$–1024). Yet, as discussed in the manuscript, the `tlrmvnmvt` and the `TruncatedNormal` packages may lead to probability estimates equaling zero, rendering them impractical and leaving expectation propagation to be the only viable option among the considered competitors, due to the high computational cost of Ridgway's sequential Monte Carlo approach in high-dimensional settings. This is particularly relevant when interest is on tail probabilities, which easily appear in high dimensions. Numerical experiments showed that these underflow issues were encountered even increasing the number of samples to 10^5 or higher, thus these problems cannot be solved by simply increasing the number of samples, creating an open problem in the literature, which can be addressed via the proposed expectation propagation approach. In the experiments, we fixed a conservative convergence tolerance on relative changes of the expectation propagation parameters, as we focused on showing that expectation propagation can compute Gaussian cumulative distribution functions with high precision, but lower computational costs could be obtained by using a higher tolerance which would decrease the number of expectation propagation iterations.

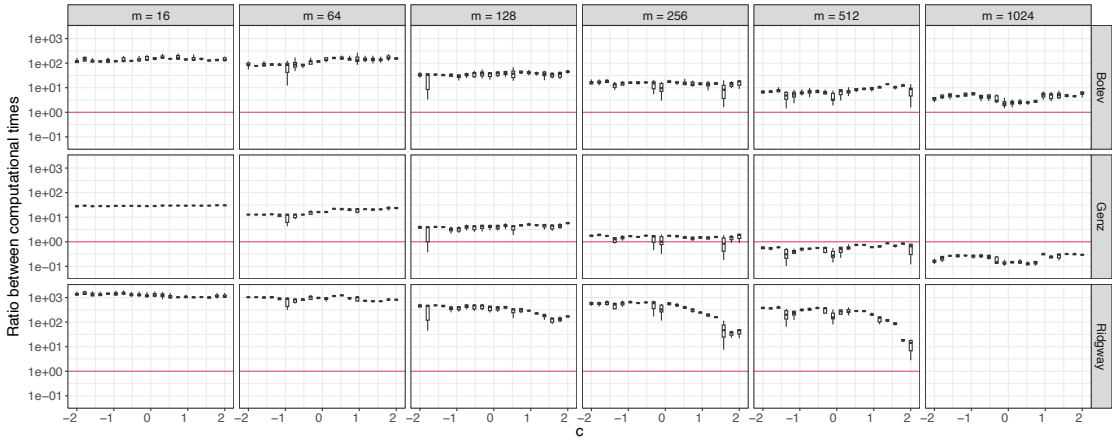

Fig. S.12: Boxplots of the ratios between the computational times obtained by the sampling-based methods (Botev (2017), Genz (1992), and Ridgway (2016)) and the expectation propagation algorithm with correlation matrix generated under case (i). Results for the methods of Ridgway (2016) in dimension $m = 1024$ were not computed due to their high computational cost.

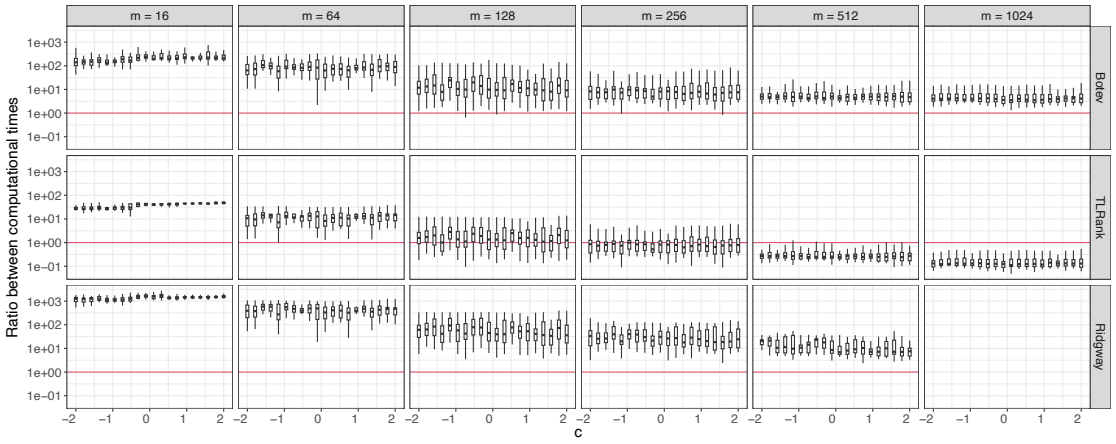

Fig. S.13: Boxplots of the ratios between the computational times obtained by the sampling-based methods (Botev (2017), Cao et al. (2021), and Ridgway (2016)) and expectation propagation algorithm with covariance matrix with correlation matrix generated under case (ii). The boxplots contain the results for all the values of ρ considered. Results for the methods of Ridgway (2016) in dimension $m = 1024$ were not computed due to their high computational cost.

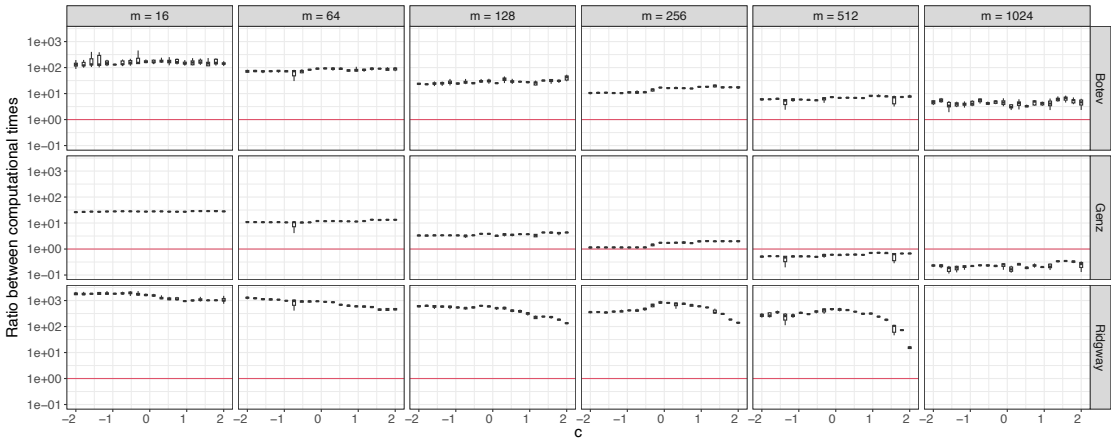

Fig. S.14: Boxplots of the ratios between the computational times obtained by the sampling-based methods (Botev (2017), Genz (1992), and Ridgway (2016)) and the expectation propagation algorithm with correlation matrix generated under case (iii). Results for the methods of Ridgway (2016) in dimension $m = 1024$ were not computed due to their high computational cost.

Lastly, to investigate the trade-off between computational costs and accuracy, we compared the estimate given by expectation propagation with that of Genz's, Botev's and Ridgway's sampling methods for varying numbers of samples $N \in \{5000, 10000, 20000, 50000\}$ in the three considered cases for the form of the matrix Σ (for case (ii), we set $\rho = 0.5$). We focused on $u = 0$ and $m = 256$ to avoid possible underflow or computational issues associated to some sampling approaches. The results are reported in Figures S.15 and S.16. It emerges that for cases (i) and (iii), expectation propagation gives virtually the same estimates as Botev's method with 50000 samples, at a fraction of the computational time, showing an extreme accuracy which makes it

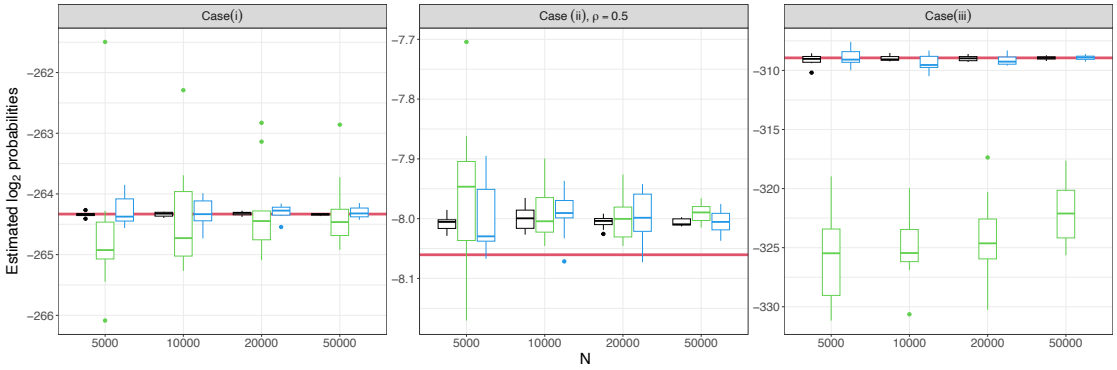

Fig. S.15: Boxplots of the estimates of $\log_2 \Phi_m(u;\Sigma)$ obtained by the sampling-based methods: Botev (2017) (black), Genz (1992) (green), and Ridgway (2016) (blue), across 10 replications. The horizontal red line represent the value obtained by the deterministic estimation of the proposed expectation propagation. All methods are evaluated at $u = 0$ with $m = 256$, varying the sample size N. Each panel corresponds to a different scenario for generating the correlation matrix Σ.

even more precise than the estimates obtained with the other sampling techniques with 50000, which are not immune to biases when small probabilities are considered: see for instance the bias of Genz's method for case (iii). On the other hand, case (ii) appears to be slightly more problematic for expectation propagation, as there seems to be some bias in the estimates, which nevertheless reduces when tail probabilities are taken into account. This suggests that expectation propagation alone might not have the same accuracy on matrices with a low-rank structure, where instead the tile low-rank method proposed in Cao et al. (2021) finds its ideal setting. Thus, this may motivate future research about theoretical guarantees for expectation propagation estimates or on even more accurate methods for the estimation of the marginal likelihood, like, e.g., estimates based on non-symmetric approximations or a combination of expectation propagation with importance sampling, where the Gaussian posterior approximation returned by expectation propagation is used as importance density in the computation of the marginal likelihood of the dual probit model of Proposition 1, highlighting the importance and the broad applicability of the main result of the manuscript. This method would be more demanding from a computational point of view, requiring an additional sampling step from m-variate random Gaussians, but would benefit from theoretical guarantees of unbiasedness inherited from importance sampling.

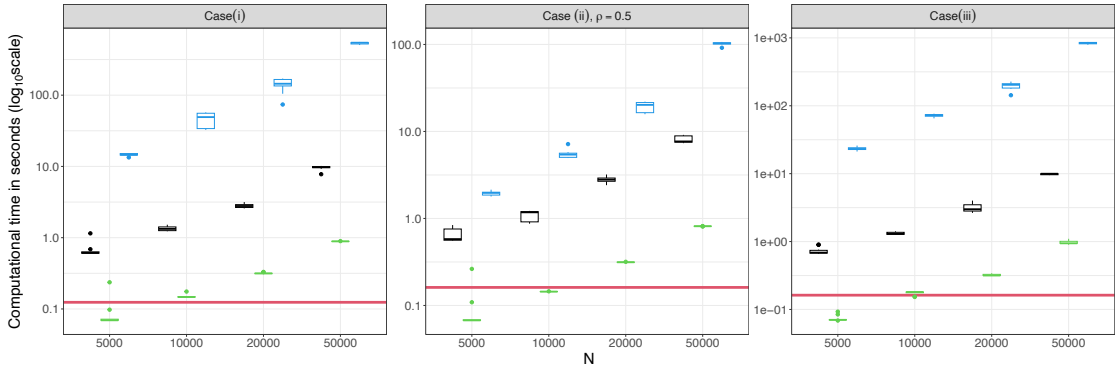

Fig. S.16: Boxplots of the computational time, in seconds, needed by the sampling-based methods: Botev (2017) (black), Genz (1992) (green), and Ridgway (2016) (blue), across 10 replications. The horizontal red line indicates the computational time, in seconds, of the proposed expectation propagation method. All methods are evaluated at $u = 0$ with $m = 256$, varying the sample size N. Each panel corresponds to a different scenario for generating the correlation matrix Σ.

REFERENCES

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

ANCESCHI, N., FASANO, A., FRANZOLINI, B. & REBAUDO, G. (2024). Scalable expectation propagation for generalized linear models. *arXiv preprint arXiv:2407.02128* .

AZZALINI, A. (2014). *The skew-normal and related families*, vol. 3. Cambridge University Press.

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

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

DAVIES, P. I. & HIGHAM, N. J. (2000). Numerically stable generation of correlation matrices and their factors. *Bit Numerical Mathematics* **40**, 640–651.

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

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

MINKA, T. P. (2001a). Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, USA: Morgan Kaufmann Publishers Inc.

MINKA, T. P. (2001b). *A Family of Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, Massachusetts Institute of Technology.

PETERSEN, K. B., PEDERSEN, M. S. et al. (2008). The matrix cookbook. *Technical University of Denmark* **7**, 510.

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

SEEGER, M. (2008). Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research* **9**, 759–813.

SEEGER, M., GERWINN, S. & BETHGE, M. (2007). Bayesian inference for sparse generalized linear models. In *Machine Learning: ECML 2007*.

TONG, Y. L. (1990). *The multivariate normal distribution*. Springer New York.

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

ZHOU, J., ORMEROD, J. T. & GRAZIAN, C. (2023). Fast expectation propagation for heteroscedastic, lasso-penalized, and quantile regression. *Journal of Machine Learning Research* **24**, 1–39.