Efficient computation of predictive probabilities in probit models via Expectation Propagation

Augusto Fasano[1], Niccolò Anceschi[2], Beatrice Franzolini[3]
and Giovanni Rebaudo[1,4]

[1] Collegio Carlo Alberto, Turin, IT (augusto.fasano@carloalberto.org)
[2] Duke University, Durham, USA (niccolo.anceschi@duke.edu)
[3] A*STAR, Singapore, SG (beatricef@sics.a-star.edu.sg)
[4] University of Turin, Turin, IT (giovanni.rebaudo@unito.it)

Abstract: Binary regression models represent a popular model-based approach for binary classification. In the Bayesian framework, computational challenges in the form of the posterior distribution motivate still-ongoing fruitful research. Here, we focus on the computation of posterior predictive probabilities in Bayesian probit models via Expectation Propagation (EP). Leveraging more general results in recent literature, we show that such predictive probabilities admit a closed-form expression. Improvements over state-of-the-art approaches are shown in a simulation study.

Keywords: probit model, expectation propagation, Bayesian inference, extended multivariate skew-normal distribution

1 Introduction

Binary regression models represent a default model-based approach for binary classification. Although the theory in the frequentist setting is well established, flourishing research is still ongoing in the Bayesian framework, where such models are also used as benchmarks for posterior computations (Chopin & Ridgway, 2017). Here, we focus on the approximation of posterior predictive probabilities via Expectation Propagation (EP) in the Bayesian probit model

$$y_i \mid \boldsymbol{\beta} \overset{ind}{\sim} \textsc{Bern}\left(\Phi\left(\mathbf{x}_i^\intercal \boldsymbol{\beta}\right)\right), i = 1, \dots, n; \quad \boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}, \nu^2 \mathbf{I}_p), \tag{1}$$

with $\boldsymbol{\beta} \in \mathbb{R}^p$ the unknown vector of parameters, $\mathbf{x}_i \in \mathbb{R}^p$ the covariate vector associated with observation i and $\mathbf{I}_p$ the identity matrix of dimension p. $\Phi(t)$ denotes the cumulative distribution function of a standard Gaussian random variable evaluated at t and $\phi_p(\mathbf{t}, \mathbf{S})$ will denote the density of a p-variate Gaussian random variable with mean $\mathbf{0}$ and covariance matrix $\mathbf{S}$, evaluated at $\mathbf{t}$.

We show that the EP approximate predictive probabilities admit a closed-form expression in terms of the output parameters returned by the EP routine. Such parameters can be obtained at per-iteration cost of $O(pn \cdot \min\{p,n\})$, as shown in Anceschi *et al.* (2023) for a broad class of models and derived in full details for the probit model in Fasano *et al.* (2023).

2 Expectation Propagation (EP) review

Adapting more general results derived in Anceschi *et al.* (2023), Fasano *et al.* (2023) showed that, calling $\mathbf{y} = (y_1, \dots, y_n)$, the EP approximation $q(\boldsymbol{\beta}) \propto \prod_{i=0}^{n} q_i(\boldsymbol{\beta})$ of the posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y})$ for model (1) can be obtained by leveraging on extended skew-normal (SN) distributions (Azzalini & Capitanio, 2014). Except for $q_0(\boldsymbol{\beta})$, which is fixed equal to the prior $p(\boldsymbol{\beta})$, we take $q_i(\boldsymbol{\beta}) = \phi_p\left(\boldsymbol{\beta} - \mathbf{Q}_i^{-1}\mathbf{r}_i, \mathbf{Q}_i^{-1}\right)$, $i = 1, \dots, n$, with the optimal $\mathbf{r}_i$'s and $\mathbf{Q}_i$'s to be obtained via the EP routine. Consequently, calling $\mathbf{r}_0 = \mathbf{0}$ and $\mathbf{Q}_0 = \nu^{-2}\mathbf{I}_p$, one gets $q(\boldsymbol{\beta}) = \phi_p(\boldsymbol{\beta} - \mathbf{Q}^{-1}\mathbf{r}, \mathbf{Q}^{-1})$, with $\mathbf{r} = \sum_{i=0}^{n} \mathbf{r}_i$, $\mathbf{Q} = \sum_{i=0}^{n} \mathbf{Q}_i$. At each EP cycle, the parameters $\mathbf{r}_i$ and $\mathbf{Q}_i$ of each site $i = 1, \dots, n$ are updated by imposing that the first two moments of the global approximation $q(\boldsymbol{\beta})$ match the ones of the hybrid distribution

$$h_i(\boldsymbol{\beta}) \propto p(y_i \mid \boldsymbol{\beta}) \prod_{j \neq i} q_j(\boldsymbol{\beta}) = \Phi((2y_i - 1)\mathbf{x}_i^\intercal \boldsymbol{\beta}) \prod_{j \neq i} q_j(\boldsymbol{\beta}). \tag{2}$$

This is immediate after noticing that (2) coincides with the kernel of a multivariate extended skew-normal distribution $\text{SN}_p(\boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\alpha}_i, \tau_i)$, with

$$\boldsymbol{\xi}_i = \mathbf{Q}_{-i}^{-1}\mathbf{r}_{-i}, \;\; \boldsymbol{\Omega}_i = \mathbf{Q}_{-i}^{-1}, \;\; \boldsymbol{\alpha}_i = (2y_i - 1)\boldsymbol{\omega}_i \mathbf{x}_i, \;\; \tau_i = (2y_i - 1)(1 + \mathbf{x}_i^\intercal \boldsymbol{\Omega}_i \mathbf{x}_i)^{-1/2} \mathbf{x}_i^\intercal \boldsymbol{\xi}_i,$$

where $\mathbf{Q}_{-i} = \sum_{j \neq i} \mathbf{Q}_j$, $\mathbf{r}_{-i} = \sum_{j \neq i} \mathbf{r}_j$ and $\boldsymbol{\omega}_i = [\text{diag}(\boldsymbol{\Omega}_i)]^{1/2}$. Combining this with Woodbury's identity, Fasano *et al.* (2023) show that, for $i = 1 \dots, n$, the updated quantities $\mathbf{r}_i^{\text{NEW}}$ and $\mathbf{Q}_i^{\text{NEW}}$ equal $m_i \mathbf{x}_i$ and $k_i \mathbf{x}_i \mathbf{x}_i^\intercal$, respectively, with $k_i = -\zeta_2(\tau_i) / \left(1 + \mathbf{x}_i^\intercal \boldsymbol{\Omega}_i \mathbf{x}_i + \zeta_2(\tau_i)\mathbf{x}_i^\intercal \boldsymbol{\Omega}_i \mathbf{x}_i\right)$ and $m_i = \zeta_1(\tau_i)s_i + k_i(\boldsymbol{\Omega}_i \mathbf{x}_i)^\intercal \mathbf{r}_{-i} + k_i \zeta_1(\tau_i) s_i \mathbf{x}_i^\intercal \boldsymbol{\Omega}_i \mathbf{x}_i$, having defined $\zeta_1(x) = \phi(x)/\Phi(x)$, $\zeta_2(x) = -\zeta_1(x)^2 - x\zeta_1(x)$ and $s_i = (2y_i - 1)(1 + \mathbf{x}_i^\intercal \boldsymbol{\Omega}_i \mathbf{x}_i)^{-1/2}$. These results, combined with the efficient computation of $\boldsymbol{\Omega}_i$ and update of the covariance matrix $\mathbf{Q}^{-1}$ of the Gaussian approximation $q(\boldsymbol{\beta})$, lead to an implementation of EP having a cost per iteration $O(p^2 n)$. When p is large, and especially when $p > n$, EP can be implemented at $O(pn^2)$ cost per iteration by storing and updating only the p-dimensional vectors $\mathbf{w}_i = \boldsymbol{\Omega}_i \mathbf{x}_i = \mathbf{Q}_{-i}^{-1}\mathbf{x}_i$ and $\mathbf{v}_i = \mathbf{Q}^{-1}\mathbf{x}_i$, $i = 1, \dots, n$. Eventually, one can compute the full EP covariance matrix as

$$\mathbf{Q}^{-1} = \nu^2 \mathbf{I}_p - \nu^2 \mathbf{VKX}, \tag{3}$$

where $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_n]$, $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^\intercal$ and $\mathbf{K} = \text{diag}(k_1, \dots, k_n)$.

3 Closed-form EP posterior predictive probabilities

One of the advantages of the Gaussian approximation provided by EP is that it results in a simple closed-form expression for the approximate posterior predictive probability of observing $y_{\text{NEW}} = 1$ for new statistical units having covariate vector $\mathbf{x}_{\text{NEW}}$, namely $\text{Pr}_{\text{EP}}[y_{\text{NEW}} = 1 \mid \mathbf{y}]$. Indeed, calling $\boldsymbol{\xi}_{\text{EP}} = \mathbf{Q}^{-1}\mathbf{r}$ and $\boldsymbol{\Omega}_{\text{EP}} = \mathbf{Q}^{-1}$ so that $q(\boldsymbol{\beta}) = \phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{EP}}, \boldsymbol{\Omega}_{\text{EP}})$, it holds

$$\text{Pr}_{\text{EP}}[y_{\text{NEW}} = 1 \mid \mathbf{y}] = \mathbb{E}_{q(\boldsymbol{\beta})}\left[\Phi\left(\mathbf{x}_{\text{NEW}}^{\intercal}\boldsymbol{\beta}\right)\right] = \Phi\left(\left(1+u\right)^{-1/2}\mathbf{x}_{\text{NEW}}^{\intercal}\boldsymbol{\xi}_{\text{EP}}\right), \qquad (4)$$

where $u = \mathbf{x}_{\text{NEW}}^{\intercal}\boldsymbol{\Omega}_{\text{EP}}\mathbf{x}_{\text{NEW}}$ and the last equality follows by Lemma 7.1 in Azzalini & Capitanio (2014). The only computationally relevant part in (4) is the computation of the quadratic form u. However, when $p < n$, $\boldsymbol{\Omega}_{\text{EP}}$ is directly returned by the algorithm, and u can be computed at cost $O(p^2)$. On the other hand, when $p > n$ (or in general when p is large), this direct computation can be avoided since, by (3), $u = \nu^2\left[\mathbf{x}_{\text{NEW}}^{\intercal}\mathbf{x}_{\text{NEW}} - \left(\mathbf{V}^{\intercal}\mathbf{x}_{\text{NEW}}\right)^{\intercal}\mathbf{K}\left(\mathbf{X}\mathbf{x}_{\text{NEW}}\right)\right]$, computable at cost $O(pn)$. Thus, Equation (4) provides an efficient closed-form approximation of the exact posterior predictive probability $\text{Pr}[y_{\text{NEW}} = 1 \mid \mathbf{y}]$, which can be computed at cost $O(p \cdot \min\{p, n\})$ from the EP parameters.

4 Simulation study

We show with a simulation study the advantages of combining the efficient EP implementation presented in Fasano *et al.* (2023) with the efficient computation of the posterior predictive probabilities presented in Section 3. Fixing $n = 100$ and $\nu^2 = 25$, we compute the posterior predictive probabilities for $\tilde{n} = 50$ test units in five different scenarios with synthetic data, for $p = 50, 100, 200, 400$ and 800. We compare the approximate posterior predictive probabilities obtained with EP and with the partially-factorized variational approximation (PFM-VB) (Equation (9) in Fasano *et al.* (2022)) with the ones arising from a Monte Carlo approximation exploiting i.i.d. samples from the posterior (Durante, 2019). Figure 1 shows that EP can achieve superior accuracy for $p < 2n$, while in the other settings they provide comparable results. The EP running time ranges from 0.02 to 0.12 seconds, while for PFM-VB it ranges from 0.13 to 0.23. The slightly higher cost of PFM-VB is because, after convergence, the computation of predictive probabilities requires a sampling step that takes approximately 0.12 seconds. To conclude, the results presented in this work make the computation of EP approximate posterior predictive probabilities feasible in settings where currently-available implementations are computationally impractical. Considering $p = 800$ for illustration,

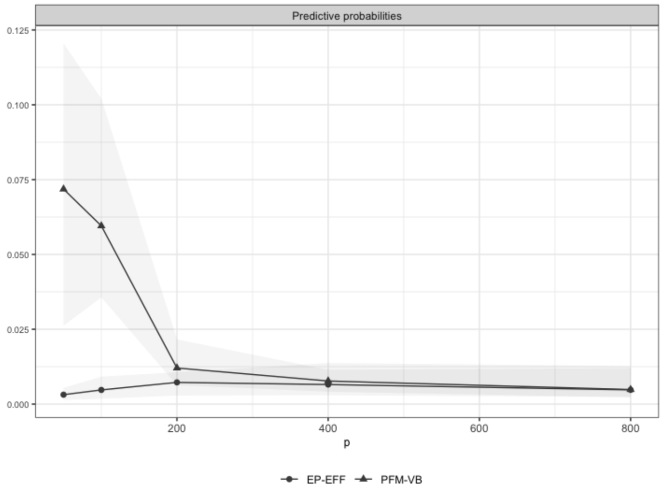

Figure 1. *For varying p, median absolute difference between the $\tilde{n} = 50$ posterior predictive probabilities resulting from* 2000 *i.i.d. samples and the ones arising from* EP *and* PFM-VB *for probit regression with $n = 100$ and $\nu^2 = 25$. Grey areas denote the first and third quartiles.*

the function `EPprobit` from the `R` package `EPGLM`, requires 140 seconds, about 1000 times slower than the efficient implementation presented here.

References

ANCESCHI, N., FASANO, A., DURANTE, D., & ZANELLA, G. 2023. Bayesian conjugacy in probit, tobit, multinomial probit and extensions: a review and new results. *J. Am. Stat. Assoc.*, **in press**.

AZZALINI, A., & CAPITANIO, A. 2014. *The skew-normal and related families*. Cambridge Univ. Press.

CHOPIN, N., & RIDGWAY, J. 2017. Leave Pima Indians alone: binary regression as a benchmark for Bayesian computation. *Stat. Sci.*, **32**, 64–87.

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

FASANO, A., DURANTE, D., & ZANELLA, G. 2022. Scalable and accurate variational Bayes for high-dimensional binary regression models. *Biometrika*, **109**, 901–919.

FASANO, A., ANCESCHI, N., FRANZOLINI, B., & REBAUDO, G. 2023. Efficient expectation propagation for high-dimensional probit models. *Preprint at http://giovannirebaudo.github.io/publications*.