The Role of Skewed Distributions in Bayesian Inference

(conjugacy, scalable approximations and asymptotics)

O'Bayes 2022: Objective Bayes Methodology Conference

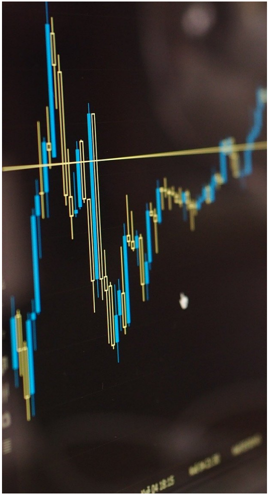
07.09.2022

Daniele Durante

daniele.durante@unibocconi.it

Regression . . .

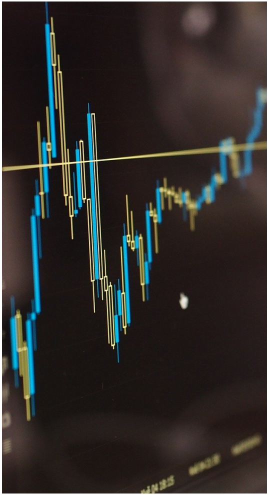
"Statisticians are engaged in an exhausting but exhilarating struggle with the biggest challenge: how to translate information into knowledge" [S. Senn]



SOURCE: https://pixabay.com/it/

Regression . . .

"Statisticians are engaged in an exhausting but exhilarating struggle with the biggest challenge: how to translate information into knowledge" [S. Senn]

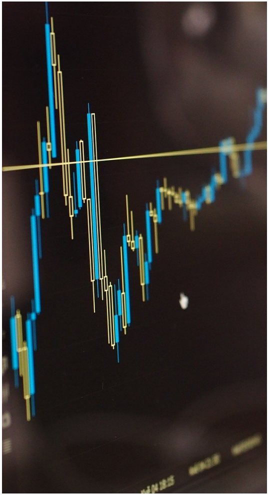


SOURCE: https://pixabay.com/it/

Regression, when possibile, is a great method to learn how the distribution of a response **y** [or functionals of it], changes with covariates.

Regression . . .

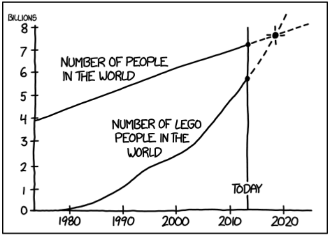
"Statisticians are engaged in an exhausting but exhilarating struggle with the biggest challenge: how to translate information into knowledge" [S. Senn]



SOURCE: https://pixabay.com/it/

Regression, when possibile, is a great method to learn how the distribution of a response **y** [or functionals of it], changes with covariates.

However, going beyond regression for Gaussian responses [either from a frequentist or Bayesian perspective], introduces some issues.



BY 2019, HUMANS WILL BE OUTNUMBERED.

SOURCE: https://xkcd.com/1281/

Bayesian probit regression

Goal: Given [conditionally] independent binary data $y_1, \ldots, y_n$ from a probit model $(y_i \mid \beta) \sim \text{Bern}[\Phi(\mathbf{x}_i^\intercal \beta)]$, $i = 1, \ldots, n$ with [in general] Gaussian prior $\beta \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$ for β, **provide inference on the posterior** $p(\beta \mid \mathbf{y})$.

Bayesian probit regression

Goal: Given [conditionally] independent binary data $y_1, \dots, y_n$ from a probit model $(y_i \mid \beta) \sim \text{Bern}[\Phi(\mathbf{x}_i^\intercal \beta)]$, $i = 1, \dots, n$ with [in general] Gaussian prior $\beta \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$ for β, **provide inference on the posterior** $p(\beta \mid \mathbf{y})$.

Applying Bayes rule, the answer to the above question is

$$p(\beta \mid \mathbf{y}) = \frac{\phi_p(\beta - \boldsymbol{\xi}; \boldsymbol{\Omega}) \prod_{i=1}^n \Phi(\mathbf{x}_i^\intercal \beta)^{y_i}[1 - \Phi(\mathbf{x}_i^\intercal \beta)]^{1-y_i}}{\int_{\mathbb{R}^p} \phi_p(\beta - \boldsymbol{\xi}; \boldsymbol{\Omega}) \prod_{i=1}^n \Phi(\mathbf{x}_i^\intercal \beta)^{y_i}[1 - \Phi(\mathbf{x}_i^\intercal \beta)]^{1-y_i} \mathrm{d}\beta}.$$

Bayesian probit regression

Goal: Given [conditionally] independent binary data $y_1, \ldots, y_n$ from a probit model $(y_i \mid \beta) \sim \text{Bern}[\Phi(\mathbf{x}_i^\intercal \beta)]$, $i = 1, \ldots, n$ with [in general] Gaussian prior $\beta \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$ for β, **provide inference on the posterior** $p(\beta \mid \mathbf{y})$.

Applying Bayes rule, the answer to the above question is

$$p(\beta \mid \mathbf{y}) = \frac{\phi_p(\beta - \boldsymbol{\xi}; \boldsymbol{\Omega}) \prod_{i=1}^n \Phi(\mathbf{x}_i^\intercal \beta)^{y_i} [1 - \Phi(\mathbf{x}_i^\intercal \beta)]^{1-y_i}}{\int_{\mathbb{R}^p} \phi_p(\beta - \boldsymbol{\xi}; \boldsymbol{\Omega}) \prod_{i=1}^n \Phi(\mathbf{x}_i^\intercal \beta)^{y_i} [1 - \Phi(\mathbf{x}_i^\intercal \beta)]^{1-y_i} \mathrm{d}\beta}.$$

However $p(\beta \mid \mathbf{y})$ does not seem to belong to some known class of distributions and the normalizing constant apparently does not have an explicit form.

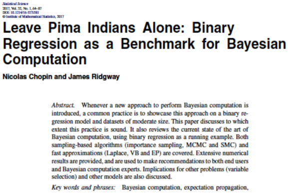
Bayesian probit regression

Goal: Given [conditionally] independent binary data $y_1, \dots, y_n$ from a probit model $(y_i \mid \beta) \sim \text{Bern}[\Phi(\mathbf{x}_i^\intercal \beta)]$, $i = 1, \dots, n$ with [in general] Gaussian prior $\beta \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$ for β, **provide inference on the posterior** $p(\beta \mid \mathbf{y})$.

Applying Bayes rule, the answer to the above question is

$$p(\beta \mid \mathbf{y}) = \frac{\phi_p(\beta - \boldsymbol{\xi}; \boldsymbol{\Omega}) \prod_{i=1}^n \Phi(\mathbf{x}_i^\intercal \beta)^{y_i} [1 - \Phi(\mathbf{x}_i^\intercal \beta)]^{1-y_i}}{\int_{\mathbb{R}^p} \phi_p(\beta - \boldsymbol{\xi}; \boldsymbol{\Omega}) \prod_{i=1}^n \Phi(\mathbf{x}_i^\intercal \beta)^{y_i} [1 - \Phi(\mathbf{x}_i^\intercal \beta)]^{1-y_i} \mathrm{d}\beta}.$$

However $p(\beta \mid \mathbf{y})$ does not seem to belong to some known class of distributions and the normalizing constant apparently does not have an explicit form.



Statistical Science
2017, Vol. 32, No. 1, 64–87
DOI: 10.1214/16-STS581
© Institute of Mathematical Statistics, 2017

Leave Pima Indians Alone: Binary Regression as a Benchmark for Bayesian Computation

Nicolas Chopin and James Ridgway

Abstract. Whenever a new approach to perform Bayesian computation is introduced, a common practice is to showcase this approach on a binary regression model and datasets of moderate size. This paper discusses to which extent this practice is sound. It also reviews the current state of the art of Bayesian computation, using binary regression as a running example. Both sampling-based algorithms (importance sampling, MCMC and SMC) and fast approximations (Laplace, VB and EP) are covered. Extensive numerical results are provided, and are used to make recommendations to both end users and Bayesian computation experts. Implications for other problems (variable selection) and other models are also discussed.

Key words and phrases: Bayesian computation, expectation propagation,

Solutions: This has motivated several methods for **Bayesian inference in probit models**, covering MCMC routines [Metropolis–Hastings, Gibbs Sampling, Hamiltonian Monte Carlo] and approximations of the posterior [Laplace, Variational Bayes, Expectation Propagation].

Unified skew–normal distribution

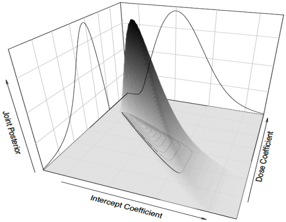
Arellano–Valle and Azzalini (2006), *Scandinavian Journal of Statistics*

Previous methods are still sub–optimal compared to cases in which the posterior belongs to a known and tractable class. **This could allow analytical posterior inference for Bayesian probit regression**. Indeed, the posterior is a SUN.

Unified skew–normal distribution

Arellano–Valle and Azzalini (2006), *Scandinavian Journal of Statistics*

Previous methods are still sub–optimal compared to cases in which the posterior belongs to a known and tractable class. **This could allow analytical posterior inference for Bayesian probit regression**. Indeed, the posterior is a SUN.



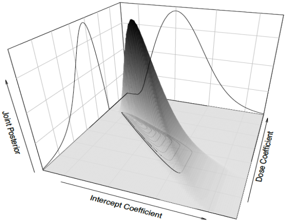
Unified skew–normal random variable (SUN)

Generalizes the multivariate SN, $\boldsymbol{\beta} \sim \mathrm{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ whose density $2\phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega})\Phi[\boldsymbol{\alpha}^\intercal \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi})]$ is obtained by modifying a $\mathrm{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$, with the cdf of the $\mathrm{N}(0, 1)$ evaluated at $\boldsymbol{\alpha}^\intercal \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi})$, with $\boldsymbol{\omega}$ the diagonal matrix of standard deviations from $\boldsymbol{\Omega}$. It unifies also other versions.

Unified skew–normal distribution

Arellano–Valle and Azzalini (2006), *Scandinavian Journal of Statistics*

Previous methods are still sub–optimal compared to cases in which the posterior belongs to a known and tractable class. **This could allow analytical posterior inference for Bayesian probit regression**. Indeed, the posterior is a SUN.



Unified skew–normal random variable (SUN)

Generalizes the multivariate SN, $\boldsymbol{\beta} \sim \mathrm{SN}_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ whose density $2\phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega})\Phi[\boldsymbol{\alpha}^\intercal \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi})]$ is obtained by modifying a $\mathrm{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$, with the cdf of the $\mathrm{N}(0,1)$ evaluated at $\boldsymbol{\alpha}^\intercal \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi})$, with $\boldsymbol{\omega}$ the diagonal matrix of standard deviations from $\boldsymbol{\Omega}$. It unifies also other versions.

More precisely, if $\boldsymbol{\beta} \sim \mathrm{SUN}_{p,q}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, with $\boldsymbol{\xi} \in \mathbb{R}^p$, $\boldsymbol{\Delta} \in \mathbb{R}^{p \times q}$, $\boldsymbol{\gamma} \in \mathbb{R}^q$ and $\boldsymbol{\Omega}^*$—having block entries $\boldsymbol{\Omega}^*_{[11]} = \boldsymbol{\Gamma}$, $\boldsymbol{\Omega}^*_{[22]} = \bar{\boldsymbol{\Omega}}$ and $\boldsymbol{\Omega}^*_{[21]} = \boldsymbol{\Omega}^{*\intercal}_{[12]} = \boldsymbol{\Delta}$—a full–rank correlation matrix, then the density is

$$\phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \frac{\Phi_q(\boldsymbol{\gamma} + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta})}{\Phi_q(\boldsymbol{\gamma}; \boldsymbol{\Gamma})}, \qquad (1)$$

Unified skew–normal conjugacy in probit regression

Durante (2019), *Biometrika*

The posterior distribution $p(\beta \mid \mathbf{y})$ for the coefficients of a probit regression $(y_i \mid \beta) \sim \text{Bern}[\Phi(\mathbf{x}_i^{\intercal}\beta)]$, $i = 1, \ldots, n$, coincides with a **unified skew–normal** (SUN) [Arellano–Valle and Azzalini, 2006], under Gaussian priors $\beta \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$.

Unified skew–normal conjugacy in probit regression

Durante (2019), *Biometrika*

The posterior distribution $p(\beta \mid \mathbf{y})$ for the coefficients of a probit regression $(y_i \mid \beta) \sim \text{Bern}[\Phi(\mathbf{x}_i^\intercal \beta)]$, $i = 1, \ldots, n$, coincides with a **unified skew–normal** ($\textsc{sun}$) [Arellano–Valle and Azzalini, 2006], under Gaussian priors $\beta \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$.

Main Theorem. If $(y_i \mid \beta) \sim \text{Bern}[\Phi(\mathbf{x}_i^\intercal \beta)]$, $i = 1, \ldots, n$ and $\beta \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$:

$$(\beta \mid \mathbf{y}) \sim \textsc{sun}_{p,n}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \bar{\boldsymbol{\Omega}}\boldsymbol{\omega}\mathbf{D}^\intercal \mathbf{s}^{-1}, \mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}, \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1}),$$

for every $\mathbf{D} = \text{diag}(2y_1 - 1, \ldots, 2y_n - 1)\mathbf{X} \in \mathbb{R}^{n \times p}$ and any $n \times n$ positive diagonal matrix of standard deviations $\mathbf{s} = [(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n) \odot \mathbf{I}_n]^{1/2}$.

Unified skew–normal conjugacy in probit regression

Durante (2019), *Biometrika*

The posterior distribution $p(\beta \mid \mathbf{y})$ for the coefficients of a probit regression $(y_i \mid \beta) \sim \text{Bern}[\Phi(\mathbf{x}_i^\intercal \beta)]$, $i = 1, \ldots, n$, coincides with a **unified skew–normal** (SUN) [Arellano–Valle and Azzalini, 2006], under Gaussian priors $\beta \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$.

Main Theorem. If $(y_i \mid \beta) \sim \text{Bern}[\Phi(\mathbf{x}_i^\intercal \beta)]$, $i = 1, \ldots, n$ and $\beta \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$:

$$(\beta \mid \mathbf{y}) \sim \text{SUN}_{p,n}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \bar{\boldsymbol{\Omega}} \omega \mathbf{D}^\intercal \mathbf{s}^{-1}, \mathbf{s}^{-1} \mathbf{D} \boldsymbol{\xi}, \mathbf{s}^{-1}(\mathbf{D} \boldsymbol{\Omega} \mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1}),$$

for every $\mathbf{D} = \text{diag}(2y_1 - 1, \ldots, 2y_n - 1)\mathbf{X} \in \mathbb{R}^{n \times p}$ and any $n \times n$ positive diagonal matrix of standard deviations $\mathbf{s} = [(\mathbf{D} \boldsymbol{\Omega} \mathbf{D}^\intercal + \mathbf{I}_n) \odot \mathbf{I}_n]^{1/2}$.

Sketch **proof**: Note $p(\beta \mid \mathbf{y}) \propto \phi_p(\beta - \xi; \boldsymbol{\Omega})\Phi_n(\mathbf{D}\beta; \mathbf{I}_n)$ and that the kernel of $\text{SUN}_{p,n}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \gamma, \boldsymbol{\Gamma})$ is $\phi_p(\beta - \xi; \boldsymbol{\Omega})\Phi_n(\gamma + \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \omega^{-1}(\beta - \xi); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta})$.

Unified skew–normal conjugacy in probit regression

Durante (2019), *Biometrika*

The posterior distribution $p(\beta \mid \mathbf{y})$ for the coefficients of a probit regression $(y_i \mid \beta) \sim \text{Bern}[\Phi(\mathbf{x}_i^\intercal \beta)]$, $i = 1, \ldots, n$, coincides with a **unified skew–normal** (SUN) [Arellano–Valle and Azzalini, 2006], under Gaussian priors $\beta \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$.

Main Theorem. If $(y_i \mid \beta) \sim \text{Bern}[\Phi(\mathbf{x}_i^\intercal \beta)]$, $i = 1, \ldots, n$ and $\beta \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$:

$$(\beta \mid \mathbf{y}) \sim \text{SUN}_{p,n}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \bar{\boldsymbol{\Omega}}\omega\mathbf{D}^\intercal\mathbf{s}^{-1}, \mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}, \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1}),$$

for every $\mathbf{D} = \text{diag}(2y_1 - 1, \ldots, 2y_n - 1)\mathbf{X} \in \mathbb{R}^{n \times p}$ and any $n \times n$ positive diagonal matrix of standard deviations $\mathbf{s} = [(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n) \odot \mathbf{I}_n]^{1/2}$.

Sketch **proof**: Note $p(\beta \mid \mathbf{y}) \propto \phi_p(\beta - \boldsymbol{\xi}; \boldsymbol{\Omega})\Phi_n(\mathbf{D}\beta; \mathbf{I}_n)$ and that the kernel of $\text{SUN}_{p,n}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \gamma, \boldsymbol{\Gamma})$ is $\phi_p(\beta - \boldsymbol{\xi}; \boldsymbol{\Omega})\Phi_n(\gamma + \boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\omega^{-1}(\beta - \boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta})$.

Remark: **Whole SUN class is conjugate to probit**. Moreover, SUN has (i) closure properties [inference on $(\beta_j \mid \mathbf{y})$], (ii) normalizing constant fairly easy to compute [prediction and variable selection], (iii) simple additive representation [iid sampling], (iv) explicit moment generating function [posterior moments].

Additive representation

To highlight the role of the hyperparameters $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$, along with that of the data $\mathbf{y}$ and $\mathbf{X}$, let us consider a stochastic representation of the SUN posterior.

If $(\beta \mid \mathbf{y}) \sim \mathrm{SUN}_{p,n}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \bar{\boldsymbol{\Omega}}\omega\mathbf{D}^\intercal\mathbf{s}^{-1}, \mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}, \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1})$, then

$$(\beta \mid \mathbf{y}) \stackrel{d}{=} \boldsymbol{\xi} + \omega[\mathbf{V}_0 + \bar{\boldsymbol{\Omega}}\omega\mathbf{D}^\intercal(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)^{-1}\mathbf{s}\mathbf{V}_1], \quad (\mathbf{V}_0 \perp \mathbf{V}_1) \tag{2}$$

with $\mathbf{V}_0 \sim \mathrm{N}_p(\mathbf{0}, \bar{\boldsymbol{\Omega}} - \bar{\boldsymbol{\Omega}}\omega\mathbf{D}^\intercal(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)^{-1}\mathbf{D}\omega\bar{\boldsymbol{\Omega}})$, and $\mathbf{V}_1$ from an n–variate Gaussian $\mathrm{N}_n(\mathbf{0}, \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1})$ truncated below $-\mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}$.

Additive representation

To highlight the role of the hyperparameters $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$, along with that of the data $\mathbf{y}$ and $\mathbf{X}$, let us consider a stochastic representation of the SUN posterior.

If $(\beta \mid \mathbf{y}) \sim \text{SUN}_{p,n}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \bar{\boldsymbol{\Omega}}\omega\mathbf{D}^\intercal\mathbf{s}^{-1}, \mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}, \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1})$, then

$$(\beta \mid \mathbf{y}) \stackrel{\mathrm{d}}{=} \boldsymbol{\xi} + \omega[\mathbf{V}_0 + \bar{\boldsymbol{\Omega}}\omega\mathbf{D}^\intercal(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)^{-1}\mathbf{s}\mathbf{V}_1], \quad (\mathbf{V}_0 \perp \mathbf{V}_1) \tag{2}$$

with $\mathbf{V}_0 \sim \mathrm{N}_p(\mathbf{0}, \bar{\boldsymbol{\Omega}} - \bar{\boldsymbol{\Omega}}\omega\mathbf{D}^\intercal(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)^{-1}\mathbf{D}\omega\bar{\boldsymbol{\Omega}})$, and $\mathbf{V}_1$ from an n–variate Gaussian $\mathrm{N}_n(\mathbf{0}, \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1})$ truncated below $-\mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}$.

Comments: The above representation provides some useful insights.

- $\boldsymbol{\xi}$ has a main role on location, but has also an effect in controlling departures from normality both in terms of skewness and excess kurtosis.
- $\boldsymbol{\Omega}$ has a main effect on scale and dependence, but contributes also to the shape in controlling the weight assigned to $\mathbf{V}_1$.
- Data in $\mathbf{D}$ play more than a role in location, scale and departures from normality. If $\mathbf{D} \approx \mathbf{0}$, $\mathbf{V}_1$ has a negligible importance compared to $\mathbf{V}_0$.

SUN is closed under marginalization, linear combinations and conditioning. Adapting these results to the unified skew–normal in the previous theorem, the marginal posteriors $(\beta_j \mid \mathbf{y})$, $j = 1, \dots, p$, still belong to the SUN family, and

$$\mathbb{E}(\beta \mid \mathbf{y}) = \boldsymbol{\xi} + \Phi_n(\mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}; \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1})^{-1}\boldsymbol{\Omega}\mathbf{D}^\intercal\mathbf{s}^{-1}\boldsymbol{\eta},$$

where $\boldsymbol{\eta}$ is a simple function of the SUN parameters.

SUN is closed under marginalization, linear combinations and conditioning. Adapting these results to the unified skew–normal in the previous theorem, the marginal posteriors $(\beta_j \mid \mathbf{y})$, $j = 1, \dots, p$, still belong to the SUN family, and

$$\mathbb{E}(\beta \mid \mathbf{y}) = \boldsymbol{\xi} + \Phi_n(\mathbf{s}^{-1}\mathbf{D}\xi; \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1})^{-1}\boldsymbol{\Omega}\mathbf{D}^\intercal\mathbf{s}^{-1}\boldsymbol{\eta},$$

where $\boldsymbol{\eta}$ is a simple function of the SUN parameters.

It is also possible to obtain closed–form expressions for posterior predictive probabilities $\text{pr}(y_{\text{new}} = 1 \mid \mathbf{y}) = \int \Phi(\mathbf{x}_{\text{new}}^\intercal \beta) p(\beta \mid \mathbf{y}) \text{d}\beta$ and the marginal likelihood $\int p(\mathbf{y} \mid \mathcal{M}_k, \beta_{\mathcal{J}_k}) p(\beta_{\mathcal{J}_k} \mid \mathcal{M}_k) \text{d}\beta_{\mathcal{J}_k}$ of a given model $\mathcal{M}_k$.

$$\text{pr}(y_{\text{new}} = 1 \mid \mathbf{y}) = \frac{\Phi_{n+1}(\mathbf{s}_{\text{new}}^{-1}\mathbf{D}_{\text{new}}\xi; \mathbf{s}_{\text{new}}^{-1}(\mathbf{D}_{\text{new}}\boldsymbol{\Omega}\mathbf{D}_{\text{new}}^\intercal + \mathbf{I}_{n+1})\mathbf{s}_{\text{new}}^{-1})}{\Phi_n(\mathbf{s}^{-1}\mathbf{D}\xi; \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1})}.$$

The marginal likelihood is instead $\Phi_n(\mathbf{s}_k^{-1}\mathbf{D}_k\xi_k; \mathbf{s}_k^{-1}(\mathbf{D}_k\boldsymbol{\Omega}_k\mathbf{D}_k^\intercal + \mathbf{I}_n)\mathbf{s}_k^{-1})$.

SUN is closed under marginalization, linear combinations and conditioning. Adapting these results to the unified skew–normal in the previous theorem, the marginal posteriors $(\beta_j \mid \mathbf{y})$, $j = 1, \ldots, p$, still belong to the SUN family, and

$$\mathbb{E}(\beta \mid \mathbf{y}) = \boldsymbol{\xi} + \Phi_n(\mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}; \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1})^{-1}\boldsymbol{\Omega}\mathbf{D}^\intercal\mathbf{s}^{-1}\boldsymbol{\eta},$$

where $\boldsymbol{\eta}$ is a simple function of the SUN parameters.

It is also possible to obtain closed–form expressions for posterior predictive probabilities $\text{pr}(y_{\text{new}} = 1 \mid \mathbf{y}) = \int \Phi(\mathbf{x}_{\text{new}}^\intercal \beta) p(\beta \mid \mathbf{y}) \text{d}\beta$ and the marginal likelihood $\int p(\mathbf{y} \mid \mathcal{M}_k, \beta_{\mathcal{J}_k}) p(\beta_{\mathcal{J}_k} \mid \mathcal{M}_k) \text{d}\beta_{\mathcal{J}_k}$ of a given model $\mathcal{M}_k$.

$$\text{pr}(y_{\text{new}} = 1 \mid \mathbf{y}) = \frac{\Phi_{n+1}(\mathbf{s}_{\text{new}}^{-1}\mathbf{D}_{\text{new}}\boldsymbol{\xi}; \mathbf{s}_{\text{new}}^{-1}(\mathbf{D}_{\text{new}}\boldsymbol{\Omega}\mathbf{D}_{\text{new}}^\intercal + \mathbf{I}_{n+1})\mathbf{s}_{\text{new}}^{-1})}{\Phi_n(\mathbf{s}^{-1}\mathbf{D}\boldsymbol{\xi}; \mathbf{s}^{-1}(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^\intercal + \mathbf{I}_n)\mathbf{s}^{-1})}.$$

The marginal likelihood is instead $\Phi_n(\mathbf{s}_k^{-1}\mathbf{D}_k\boldsymbol{\xi}_k; \mathbf{s}_k^{-1}(\mathbf{D}_k\boldsymbol{\Omega}_k\mathbf{D}_k^\intercal + \mathbf{I}_n)\mathbf{s}_k^{-1})$.

Problem: Inference requires sampling from n–variate truncated normals or evaluation of cumulative distribution functions $\Phi_n(\cdot)$ of n–variate Gaussians.

Closed–form filter for dynamic probit models

Fasano, Rebaudo, **Durante**, Petrone (2021), *Statistics and Computing*

Goal: **Closed–form recursive expressions** for predictive $p(\beta_t \mid \mathbf{y}_{1:t-1})$, filtering $p(\beta_t \mid \mathbf{y}_{1:t})$ and smoothing $p(\beta_{1:n} \mid \mathbf{y}_{1:n})$ distributions in the dynamic model

$$
\begin{aligned}
(y_t \mid \beta_t) &\sim \text{Bern}[\Phi(\mathbf{x}_t^\intercal \beta_t)] \rightarrow p(y_t \mid \beta_t) = \Phi[(2y_t - 1)\mathbf{x}_t^\intercal \beta_t] \\
\beta_t &= \mathbf{G}_t \beta_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{N}_p(\mathbf{0}, \mathbf{W}_t), \; t = 1 \ldots, n, \; \beta_0 \sim \text{N}_p(\mathbf{a}_0, \mathbf{P}_0)
\end{aligned}
$$

Closed–form filter for dynamic probit models

Fasano, Rebaudo, **Durante**, Petrone (2021), *Statistics and Computing*

Goal: **Closed–form recursive expressions** for predictive $p(\beta_t \mid \mathbf{y}_{1:t-1})$, filtering $p(\beta_t \mid \mathbf{y}_{1:t})$ and smoothing $p(\beta_{1:n} \mid \mathbf{y}_{1:n})$ distributions in the dynamic model

$$
\begin{aligned}
(y_t \mid \beta_t) &\sim \text{Bern}[\Phi(\mathbf{x}_t^\intercal \beta_t)] \rightarrow p(y_t \mid \beta_t) = \Phi[(2y_t - 1)\mathbf{x}_t^\intercal \beta_t] \\
\beta_t &= \mathbf{G}_t \beta_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{N}_p(\mathbf{0}, \mathbf{W}_t), \ t = 1 \ldots, n, \ \ \beta_0 \sim \text{N}_p(\mathbf{a}_0, \mathbf{P}_0)
\end{aligned}
$$

Hint: Note that $p(\beta_1 \mid y_1) \propto \phi_p(\beta_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi[(2y_1 - 1)\mathbf{x}_1^\intercal \beta_1]$.

Closed–form filter for dynamic probit models

Fasano, Rebaudo, **Durante**, Petrone (2021), *Statistics and Computing*

Goal: **Closed–form recursive expressions** for predictive $p(\beta_t \mid \mathbf{y}_{1:t-1})$, filtering $p(\beta_t \mid \mathbf{y}_{1:t})$ and smoothing $p(\beta_{1:n} \mid \mathbf{y}_{1:n})$ distributions in the dynamic model

$$
\begin{aligned}
(y_t \mid \beta_t) &\sim \text{Bern}[\Phi(\mathbf{x}_t^\intercal \beta_t)] \rightarrow p(y_t \mid \beta_t) = \Phi[(2y_t - 1)\mathbf{x}_t^\intercal \beta_t] \\
\beta_t &= \mathbf{G}_t \beta_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{N}_p(\mathbf{0}, \mathbf{W}_t), \; t = 1 \ldots, n, \quad \beta_0 \sim \text{N}_p(\mathbf{a}_0, \mathbf{P}_0)
\end{aligned}
$$

Hint: Note that $p(\beta_1 \mid y_1) \propto \phi_p(\beta_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi[(2y_1 - 1)\mathbf{x}_1^\intercal \beta_1]$.

Main Theorem [closed–form filter for probit state–space models]

1. **Filtering** $[t-1]$ $\rightarrow$ **Predictive** $[t]$: If $(\beta_{t-1} \mid \mathbf{y}_{1:t-1})$ is a $\text{SUN}_{p,t-1}$ and $\beta_t = \mathbf{G}_t \beta_{t-1} + \varepsilon_t$, with $\varepsilon_t \sim \text{N}_p(\mathbf{0}, \mathbf{W}_t)$, then $(\beta_t \mid \mathbf{y}_{1:t-1})$ is also a $\text{SUN}_{p,t-1}$ with updated parameters [closure under linear combinations].
2. **Predictive** $[t]$ $\rightarrow$ **Filtering** $[t]$: if $(\beta_t \mid \mathbf{y}_{1:t-1})$ is $\text{SUN}_{p,t-1}$ and $p(y_t \mid \beta_t)$ is a probit likelihood, then $p(\beta_t \mid \mathbf{y}_{1:t}) \propto p(\beta_t \mid \mathbf{y}_{1:t-1})\Phi[(2y_t - 1)\mathbf{x}_t^\intercal \beta_t]$ is also $\text{SUN}_{p,t}$ with updated parameters [SUN–probit conjugacy; **Durante**, 2019].

Closed–form filter for dynamic probit models

Fasano, Rebaudo, **Durante**, Petrone (2021), *Statistics and Computing*

Goal: **Closed–form recursive expressions** for predictive $p(\beta_t \mid \mathbf{y}_{1:t-1})$, filtering $p(\beta_t \mid \mathbf{y}_{1:t})$ and smoothing $p(\beta_{1:n} \mid \mathbf{y}_{1:n})$ distributions in the dynamic model

$$
\begin{aligned}
(y_t \mid \beta_t) &\sim \text{Bern}[\Phi(\mathbf{x}_t^\intercal \beta_t)] \rightarrow p(y_t \mid \beta_t) = \Phi[(2y_t - 1)\mathbf{x}_t^\intercal \beta_t] \\
\beta_t &= \mathbf{G}_t \beta_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{N}_p(\mathbf{0}, \mathbf{W}_t), \ t = 1 \ldots, n, \ \ \beta_0 \sim \text{N}_p(\mathbf{a}_0, \mathbf{P}_0)
\end{aligned}
$$

Hint: Note that $p(\beta_1 \mid y_1) \propto \phi_p(\beta_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi[(2y_1 - 1)\mathbf{x}_1^\intercal \beta_1]$.

Main Theorem [closed–form filter for probit state–space models]

1. **Filtering** $[t-1] \rightarrow$ **Predictive** $[t]$: If $(\beta_{t-1} \mid \mathbf{y}_{1:t-1})$ is a $\text{SUN}_{p,t-1}$ and $\beta_t = \mathbf{G}_t \beta_{t-1} + \varepsilon_t$, with $\varepsilon_t \sim \text{N}_p(\mathbf{0}, \mathbf{W}_t)$, then $(\beta_t \mid \mathbf{y}_{1:t-1})$ is also a $\text{SUN}_{p,t-1}$ with updated parameters [closure under linear combinations].
2. **Predictive** $[t] \rightarrow$ **Filtering** $[t]$: if $(\beta_t \mid \mathbf{y}_{1:t-1})$ is $\text{SUN}_{p,t-1}$ and $p(y_t \mid \beta_t)$ is a probit likelihood, then $p(\beta_t \mid \mathbf{y}_{1:t}) \propto p(\beta_t \mid \mathbf{y}_{1:t-1})\Phi[(2y_t - 1)\mathbf{x}_t^\intercal \beta_t]$ is also $\text{SUN}_{p,t}$ with updated parameters [SUN–probit conjugacy; **Durante**, 2019].

Analog of the **Kalman filter** in the context of binary state–space models.

Extension to L categories [based on **Gaussian latent utilities** $u_{i1}, \ldots, u_{iL}$].

- **[Hausman and Wise, 1978]**. $\text{pr}(y_i = l \mid \beta) = \text{pr}(u_{il} > u_{ik}, \forall k \neq l)$ with $u_{il} = \mathbf{x}_{il}^\intercal \beta + \varepsilon_{il}$, $\varepsilon_i \sim \text{N}_L(\mathbf{0}, \mathbf{\Sigma})$ for each $l = 1, \ldots, L$ and $i = 1, \ldots, n$.
- **[Stern, 1992]**. $\text{pr}(y_i = l \mid \beta) = \text{pr}(u_{il} > u_{ik}, \forall k \neq l)$ with $u_{il} = \mathbf{x}_i^\intercal \beta_l + \varepsilon_{il}$, $\varepsilon_i \sim \text{N}_L(\mathbf{0}, \mathbf{\Sigma})$ for each $l = 1, \ldots, L$ and $i = 1, \ldots, n$.
- **[Tutz, 1991]**. Based on a nested decision process relying on sequential binary decisions with probability $\text{pr}(y_i = l \mid y_i > l - 1, \beta) = \Phi(\mathbf{x}_i^\intercal \beta_l)$.

Extension to L categories [based on **Gaussian latent utilities** $u_{i1}, \ldots, u_{iL}$].

- **[Hausman and Wise, 1978]**. $\text{pr}(y_i = l \mid \beta) = \text{pr}(u_{il} > u_{ik}, \forall k \neq l)$ with $u_{il} = \mathbf{x}_{il}^\intercal \beta + \varepsilon_{il}$, $\varepsilon_i \sim \text{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$ for each $l = 1, \ldots, L$ and $i = 1, \ldots, n$.
- **[Stern, 1992]**. $\text{pr}(y_i = l \mid \beta) = \text{pr}(u_{il} > u_{ik}, \forall k \neq l)$ with $u_{il} = \mathbf{x}_i^\intercal \beta_l + \varepsilon_{il}$, $\varepsilon_i \sim \text{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$ for each $l = 1, \ldots, L$ and $i = 1, \ldots, n$.
- **[Tutz, 1991]**. Based on a nested decision process relying on sequential binary decisions with probability $\text{pr}(y_i = l \mid y_i > l-1, \beta) = \Phi(\mathbf{x}_i^\intercal \beta_l)$.

Goal: **Closed–form results** for $p(\beta \mid \mathbf{y})$, when β has Gaussian or SUN prior.

SUN conjugacy in multinomial probit models

Fasano and **Durante** (2022), *Journal of Machine Learning Research*

Extension to L categories [based on **Gaussian latent utilities** $u_{i1}, \ldots, u_{iL}$].

- **[Hausman and Wise, 1978]**. $\text{pr}(y_i = l \mid \beta) = \text{pr}(u_{il} > u_{ik}, \forall k \neq l)$ with $u_{il} = \mathbf{x}_{il}^{\intercal}\beta + \varepsilon_{il}$, $\boldsymbol{\varepsilon}_i \sim \text{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$ for each $l = 1, \ldots, L$ and $i = 1, \ldots, n$.
- **[Stern, 1992]**. $\text{pr}(y_i = l \mid \beta) = \text{pr}(u_{il} > u_{ik}, \forall k \neq l)$ with $u_{il} = \mathbf{x}_i^{\intercal}\beta_l + \varepsilon_{il}$, $\boldsymbol{\varepsilon}_i \sim \text{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$ for each $l = 1, \ldots, L$ and $i = 1, \ldots, n$.
- **[Tutz, 1991]**. Based on a nested decision process relying on sequential binary decisions with probability $\text{pr}(y_i = l \mid y_i > l-1, \beta) = \Phi(\mathbf{x}_i^{\intercal}\beta_l)$.

Goal: **Closed–form results** for $p(\beta \mid \mathbf{y})$, when β has Gaussian or SUN prior.

Hint: The above models admit likelihood of the form $p(\mathbf{y} \mid \beta) = \Phi_m(\bar{\mathbf{X}}\beta; \boldsymbol{\Lambda})$, where $\bar{\mathbf{X}}$ and $\boldsymbol{\Lambda}$ are suitable design and covariance matrices, respectively.

SUN conjugacy in multinomial probit models

Fasano and **Durante** (2022), *Journal of Machine Learning Research*

Extension to L categories [based on **Gaussian latent utilities** $u_{i1}, \dots, u_{iL}$].

- **[Hausman and Wise, 1978]**. $\text{pr}(y_i = l \mid \beta) = \text{pr}(u_{il} > u_{ik}, \forall k \neq l)$ with $u_{il} = \mathbf{x}_{il}^\intercal \beta + \varepsilon_{il}$, $\varepsilon_i \sim \text{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$ for each $l = 1, \dots, L$ and $i = 1, \dots, n$.
- **[Stern, 1992]**. $\text{pr}(y_i = l \mid \beta) = \text{pr}(u_{il} > u_{ik}, \forall k \neq l)$ with $u_{il} = \mathbf{x}_i^\intercal \beta_l + \varepsilon_{il}$, $\varepsilon_i \sim \text{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$ for each $l = 1, \dots, L$ and $i = 1, \dots, n$.
- **[Tutz, 1991]**. Based on a nested decision process relying on sequential binary decisions with probability $\text{pr}(y_i = l \mid y_i > l-1, \beta) = \Phi(\mathbf{x}_i^\intercal \beta_l)$.

Goal: **Closed–form results** for $p(\beta \mid \mathbf{y})$, when β has Gaussian or SUN prior.

Hint: The above models admit likelihood of the form $p(\mathbf{y} \mid \beta) = \Phi_m(\bar{\mathbf{X}}\beta; \boldsymbol{\Lambda})$, where $\bar{\mathbf{X}}$ and $\boldsymbol{\Lambda}$ are suitable design and covariance matrices, respectively.

Main Theorem. If $p(\mathbf{y} \mid \beta) = \Phi_m(\bar{\mathbf{X}}\beta; \boldsymbol{\Lambda})$ and β has SUN prior (Gaussian is a special case), then $(\beta \mid \mathbf{y}) \sim \text{SUN}_{q,m'}(\xi_{\text{POST}}, \boldsymbol{\Omega}_{\text{POST}}, \boldsymbol{\Delta}_{\text{POST}}, \gamma_{\text{POST}}, \boldsymbol{\Gamma}_{\text{POST}})$.

SUN conjugacy in multinomial probit models

Fasano and **Durante** (2022), *Journal of Machine Learning Research*

Extension to L categories [based on **Gaussian latent utilities** $u_{i1}, \ldots, u_{iL}$].

- **[Hausman and Wise, 1978]**. $\text{pr}(y_i = l \mid \beta) = \text{pr}(u_{il} > u_{ik}, \forall k \neq l)$ with $u_{il} = \mathbf{x}_{il}^{\intercal}\beta + \varepsilon_{il}$, $\varepsilon_i \sim \text{N}_L(\mathbf{0}, \mathbf{\Sigma})$ for each $l = 1, \ldots, L$ and $i = 1, \ldots, n$.
- **[Stern, 1992]**. $\text{pr}(y_i = l \mid \beta) = \text{pr}(u_{il} > u_{ik}, \forall k \neq l)$ with $u_{il} = \mathbf{x}_{i}^{\intercal}\beta_l + \varepsilon_{il}$, $\varepsilon_i \sim \text{N}_L(\mathbf{0}, \mathbf{\Sigma})$ for each $l = 1, \ldots, L$ and $i = 1, \ldots, n$.
- **[Tutz, 1991]**. Based on a nested decision process relying on sequential binary decisions with probability $\text{pr}(y_i = l \mid y_i > l-1, \beta) = \Phi(\mathbf{x}_{i}^{\intercal}\beta_l)$.

Goal: **Closed–form results** for $p(\beta \mid \mathbf{y})$, when β has Gaussian or SUN prior.

Hint: The above models admit likelihood of the form $p(\mathbf{y} \mid \beta) = \Phi_m(\bar{\mathbf{X}}\beta; \mathbf{\Lambda})$, where $\bar{\mathbf{X}}$ and $\mathbf{\Lambda}$ are suitable design and covariance matrices, respectively.

Main Theorem. If $p(\mathbf{y} \mid \beta) = \Phi_m(\bar{\mathbf{X}}\beta; \mathbf{\Lambda})$ and β has SUN prior (Gaussian is a special case), then $(\beta \mid \mathbf{y}) \sim \text{SUN}_{q,m'}(\xi_{\text{POST}}, \mathbf{\Omega}_{\text{POST}}, \mathbf{\Delta}_{\text{POST}}, \gamma_{\text{POST}}, \mathbf{\Gamma}_{\text{POST}})$.

Leverage the SUN properties also for Bayesian inference in multinomial probits.

Useful augmented–data representation

Albert and Chib (1993)

Problem. **Closed–form inference** under SUN posteriors requires to deal with multivariate truncated normals and cdfs of multivariate Gaussians whose dimension grows with the sample size $n \rightarrow$ try to **approximate the posterior**.

Useful augmented–data representation

Albert and Chib (1993)

Problem. **Closed–form inference** under SUN posteriors requires to deal with multivariate truncated normals and cdfs of multivariate Gaussians whose dimension grows with the sample size $n \rightarrow$ try to **approximate the posterior**.

Bayesian probit regression model can also be expressed as

$$y_i = \mathbb{1}(z_i > 0), \text{ with } (z_i \mid \beta) \sim \mathrm{N}(\mathbf{x}_i^\intercal \beta, 1), i = 1, \ldots, n, \text{ and } \beta \sim \mathrm{N}_p(\mathbf{0}, \nu_p^2 \mathbf{I}_p).$$

Thus, we have a **dichotomized Gaussian linear regression on latent data** z_i.

Useful augmented–data representation

Albert and Chib (1993)

Problem. **Closed–form inference** under SUN posteriors requires to deal with multivariate truncated normals and cdfs of multivariate Gaussians whose dimension grows with the sample size $n \to$ try to **approximate the posterior**.

Bayesian probit regression model can also be expressed as

$$y_i = \mathbb{1}(z_i > 0), \text{ with } (z_i \mid \beta) \sim \mathrm{N}(\mathbf{x}_i^\intercal \beta, 1), i = 1, \dots, n, \text{ and } \beta \sim \mathrm{N}_p(\mathbf{0}, \nu_p^2 \mathbf{I}_p).$$

Thus, we have a **dichotomized Gaussian linear regression on latent data** z_i.

This has been widely used in the development of MCMC and VB methods.

$$(\beta \mid \mathbf{z}, \mathbf{y}) \sim \mathrm{N}_p(\mathbf{V}\mathbf{X}^\intercal \mathbf{z}, \mathbf{V}), \quad \mathbf{V} = (\nu_p^{-2}\mathbf{I}_p + \mathbf{X}^\intercal \mathbf{X})^{-1},$$

$$(z_i \mid \beta, \mathbf{z}_{-i}, \mathbf{y}) \sim \begin{cases} \mathrm{TN}[\mathbf{x}_i^\intercal \beta, 1, (0, +\infty)], & \text{if } y_i = 1, \\ \mathrm{TN}[\mathbf{x}_i^\intercal \beta, 1, (-\infty, 0)], & \text{if } y_i = 0, \end{cases} \quad \text{for } i = 1, \dots, n,$$

These full–conditionals allow implementation of Gibbs samplers [Albert and Chib, 1993] and mean–field VB with global and local variables [Consonni and Marin, 2007].

Mean–field variational Bayes for probit models

Goal: Find a tractable approximation for the joint posterior density $p(\beta, \mathbf{z} \mid \mathbf{y})$, within the MF class of densities $\mathcal{Q}_{\text{MF}} = \{q_{\text{MF}}(\beta, \mathbf{z}) : q_{\text{MF}}(\beta, \mathbf{z}) = q_{\text{MF}}(\beta) q_{\text{MF}}(\mathbf{z})\}$

Mean–field variational Bayes for probit models

Goal: Find a tractable approximation for the joint posterior density $p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})$, within the MF class of densities $\mathcal{Q}_{\text{MF}} = \{q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) : q_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) = q_{\text{MF}}(\boldsymbol{\beta}) q_{\text{MF}}(\mathbf{z})\}$

The optimal VB solution $q^*_{\text{MF}}(\boldsymbol{\beta}) q^*_{\text{MF}}(\mathbf{z})$ within this family minimizes

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

Mean–field variational Bayes for probit models

Goal: Find a tractable approximation for the joint posterior density $p(\beta, \mathbf{z} \mid \mathbf{y})$, within the MF class of densities $\mathcal{Q}_{\mathrm{MF}} = \{q_{\mathrm{MF}}(\beta, \mathbf{z}) : q_{\mathrm{MF}}(\beta, \mathbf{z}) = q_{\mathrm{MF}}(\beta) q_{\mathrm{MF}}(\mathbf{z})\}$

The optimal VB solution $q^*_{\mathrm{MF}}(\beta) q^*_{\mathrm{MF}}(\mathbf{z})$ within this family minimizes

$$\mathrm{KL}[q_{\mathrm{MF}}(\beta, \mathbf{z}) \mid\mid p(\beta, \mathbf{z} \mid \mathbf{y})] = \mathbb{E}_{q_{\mathrm{MF}}(\beta, \mathbf{z})}[\log q_{\mathrm{MF}}(\beta, \mathbf{z})] - \mathbb{E}_{q_{\mathrm{MF}}(\beta, \mathbf{z})}[\log p(\beta, \mathbf{z} \mid \mathbf{y})].$$

In practice, we maximize $\mathrm{ELBO}[q_{\mathrm{MF}}(\beta, \mathbf{z})] = -\mathrm{KL}[q_{\mathrm{MF}}(\beta, \mathbf{z}) || p(\beta, \mathbf{z} \mid \mathbf{y})] + \mathrm{c}$ via

$$q^{(t)}_{\mathrm{MF}}(\beta) \propto \exp\{\mathbb{E}_{q^{(t-1)}_{\mathrm{MF}}(\mathbf{z})} \log[p(\beta \mid \mathbf{z}, \mathbf{y})]\}, \;\; q^{(t)}_{\mathrm{MF}}(\mathbf{z}) \propto \exp\{\mathbb{E}_{q^{(t)}_{\mathrm{MF}}(\beta)} \log[p(\mathbf{z} \mid \beta, \mathbf{y})]\},$$

that approximates $p(\beta, \mathbf{z} \mid \mathbf{y})$ via $q^*_{\mathrm{MF}}(\beta) \prod_{i=1}^n q^*_{\mathrm{MF}}(z_i)$, where $q^*_{\mathrm{MF}}(\beta)$ is a Gaussian, while $q^*_{\mathrm{MF}}(z_i)$ are univariate truncated normals.

Mean–field variational Bayes for probit models

Goal: Find a tractable approximation for the joint posterior density $p(\beta, \mathbf{z} \mid \mathbf{y})$, within the MF class of densities $\mathcal{Q}_{\mathrm{MF}} = \{q_{\mathrm{MF}}(\beta, \mathbf{z}) : q_{\mathrm{MF}}(\beta, \mathbf{z}) = q_{\mathrm{MF}}(\beta) q_{\mathrm{MF}}(\mathbf{z})\}$

The optimal VB solution $q^*_{\mathrm{MF}}(\beta) q^*_{\mathrm{MF}}(\mathbf{z})$ within this family minimizes

$$\mathrm{KL}[q_{\mathrm{MF}}(\beta, \mathbf{z}) \mid\mid p(\beta, \mathbf{z} \mid \mathbf{y})] = \mathbb{E}_{q_{\mathrm{MF}}(\beta, \mathbf{z})}[\log q_{\mathrm{MF}}(\beta, \mathbf{z})] - \mathbb{E}_{q_{\mathrm{MF}}(\beta, \mathbf{z})}[\log p(\beta, \mathbf{z} \mid \mathbf{y})].$$

In practice, we maximize $\mathrm{ELBO}[q_{\mathrm{MF}}(\beta, \mathbf{z})] = -\mathrm{KL}[q_{\mathrm{MF}}(\beta, \mathbf{z}) || p(\beta, \mathbf{z} \mid \mathbf{y})] + \mathrm{c}$ via

$$q^{(t)}_{\mathrm{MF}}(\beta) \propto \exp\{\mathbb{E}_{q^{(t-1)}_{\mathrm{MF}}(\mathbf{z})} \log[p(\beta \mid \mathbf{z}, \mathbf{y})]\}, \;\; q^{(t)}_{\mathrm{MF}}(\mathbf{z}) \propto \exp\{\mathbb{E}_{q^{(t)}_{\mathrm{MF}}(\beta)} \log[p(\mathbf{z} \mid \beta, \mathbf{y})]\},$$

that approximates $p(\beta, \mathbf{z} \mid \mathbf{y})$ via $q^*_{\mathrm{MF}}(\beta) \prod_{i=1}^{n} q^*_{\mathrm{MF}}(z_i)$, where $q^*_{\mathrm{MF}}(\beta)$ is a Gaussian, while $q^*_{\mathrm{MF}}(z_i)$ are univariate truncated normals.

However, Fasano, **Durante**, Zanella (2022+) show that

Mean–field variational Bayes for probit models

Goal: Find a tractable approximation for the joint posterior density $p(\beta, \mathbf{z} \mid \mathbf{y})$, within the MF class of densities $\mathcal{Q}_{\text{MF}} = \{q_{\text{MF}}(\beta, \mathbf{z}) : q_{\text{MF}}(\beta, \mathbf{z}) = q_{\text{MF}}(\beta) q_{\text{MF}}(\mathbf{z})\}$

The optimal VB solution $q^*_{\text{MF}}(\beta) q^*_{\text{MF}}(\mathbf{z})$ within this family minimizes

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

In practice, we maximize $\text{ELBO}[q_{\text{MF}}(\beta, \mathbf{z})] = -\text{KL}[q_{\text{MF}}(\beta, \mathbf{z}) || p(\beta, \mathbf{z} \mid \mathbf{y})] + \text{c}$ via

$$q^{(t)}_{\text{MF}}(\beta) \propto \exp\{\mathbb{E}_{q^{(t-1)}_{\text{MF}}(\mathbf{z})} \log[p(\beta \mid \mathbf{z}, \mathbf{y})]\}, \;\; q^{(t)}_{\text{MF}}(\mathbf{z}) \propto \exp\{\mathbb{E}_{q^{(t)}_{\text{MF}}(\beta)} \log[p(\mathbf{z} \mid \beta, \mathbf{y})]\},$$

that approximates $p(\beta, \mathbf{z} \mid \mathbf{y})$ via $q^*_{\text{MF}}(\beta) \prod_{i=1}^n q^*_{\text{MF}}(z_i)$, where $q^*_{\text{MF}}(\beta)$ is a Gaussian, while $q^*_{\text{MF}}(z_i)$ are univariate truncated normals.

However, Fasano, **Durante**, Zanella (2022+) show that

Theorem: Under simple assumptions, $\liminf_{p\to\infty} \text{KL}[q^*_{\text{MF}}(\beta) \mid\mid p(\beta \mid \mathbf{y})] > 0$ almost surely (a.s.). Moreover, $\nu_p^{-1} ||\mathbb{E}_{q^*_{\text{MF}}(\beta)}(\beta)|| \to 0$ (a.s.) as $p \to \infty$, where $||\cdot||$ is the Euclidean norm. On the contrary, $\nu_p^{-1} ||\mathbb{E}_{p(\beta\mid\mathbf{y})}(\beta)|| \to \text{const} \cdot \sqrt{n} > 0$ (a.s.) as $p \to \infty$, where **const** is a strictly positive constant.

Solution: **Enlarge the class** of approximating densities in a way that still allows simple optimization and inference. In particular, we consider the partially factorized family $\mathcal{Q}_{\text{PFM}} = \{q_{\text{PFM}}(\beta, \mathbf{z}) : q_{\text{PFM}}(\beta, \mathbf{z}) = q_{\text{PFM}}(\beta \mid \mathbf{z}) \prod_{i=1}^{n} q_{\text{PFM}}(z_i)\}$.

Solution: **Enlarge the class** of approximating densities in a way that still allows simple optimization and inference. In particular, we consider the partially factorized family $\mathcal{Q}_{\text{PFM}} = \{q_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z}) : q_{\text{PFM}}(\boldsymbol{\beta}, \mathbf{z}) = q_{\text{PFM}}(\boldsymbol{\beta} \mid \mathbf{z}) \prod_{i=1}^{n} q_{\text{PFM}}(z_i)\}$.

Motivation for the use of $\mathcal{Q}_{\text{PFM}}$: $q^*_{\text{MF}}(\boldsymbol{\beta}, \mathbf{z}) = q^*_{\text{MF}}(\boldsymbol{\beta}) \prod_{i=1}^{n} q^*_{\text{MF}}(z_i)$ belongs to $\mathcal{Q}_{\text{PFM}}$, and $p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y}) = p(\boldsymbol{\beta} \mid \mathbf{z}) p(\mathbf{z} \mid \mathbf{y})$ with $p(\boldsymbol{\beta} \mid \mathbf{z}) = \phi_p(\boldsymbol{\beta} - \mathbf{V}\mathbf{X}^\intercal \mathbf{z}; \mathbf{V})$ and $p(\mathbf{z} \mid \mathbf{y}) \propto \phi_n(\mathbf{z}; \mathbf{I}_n + \nu_p^2 \mathbf{X}\mathbf{X}^\intercal) \prod_{i=1}^{n} \mathbb{1}[(2y_i - 1) z_i > 0]$ [Holmes and Held, 2006].

Solution: **Enlarge the class** of approximating densities in a way that still allows simple optimization and inference. In particular, we consider the partially factorized family $\mathcal{Q}_{\text{PFM}} = \{q_{\text{PFM}}(\beta, \mathbf{z}) : q_{\text{PFM}}(\beta, \mathbf{z}) = q_{\text{PFM}}(\beta \mid \mathbf{z}) \prod_{i=1}^{n} q_{\text{PFM}}(z_i)\}$.

Motivation for the use of $\mathcal{Q}_{\text{PFM}}$: $q^*_{\text{MF}}(\beta, \mathbf{z}) = q^*_{\text{MF}}(\beta) \prod_{i=1}^{n} q^*_{\text{MF}}(z_i)$ belongs to $\mathcal{Q}_{\text{PFM}}$, and $p(\beta, \mathbf{z} \mid \mathbf{y}) = p(\beta \mid \mathbf{z}) p(\mathbf{z} \mid \mathbf{y})$ with $p(\beta \mid \mathbf{z}) = \phi_p(\beta - \mathbf{V}\mathbf{X}^\intercal \mathbf{z}; \mathbf{V})$ and $p(\mathbf{z} \mid \mathbf{y}) \propto \phi_n(\mathbf{z}; \mathbf{I}_n + \nu_p^2 \mathbf{X}\mathbf{X}^\intercal) \prod_{i=1}^{n} \mathbb{1}[(2y_i - 1) z_i > 0]$ [Holmes and Held, 2006].

Proposition. Let $q^*_{\text{PFM}}(\beta, \mathbf{z})$ and $q^*_{\text{MF}}(\beta, \mathbf{z})$ be the optimal approximations for $p(\beta, \mathbf{z} \mid \mathbf{y})$, under PFM-VB and MF-VB, respectively. Then

$$\text{KL}[q^*_{\text{PFM}}(\beta, \mathbf{z}) \,||\, p(\beta, \mathbf{z} \mid \mathbf{y})] \leq \text{KL}[q^*_{\text{MF}}(\beta, \mathbf{z}) \,||\, p(\beta, \mathbf{z} \mid \mathbf{y})].$$

Solution: **Enlarge the class** of approximating densities in a way that still allows simple optimization and inference. In particular, we consider the partially factorized family $\mathcal{Q}_{\text{PFM}} = \{q_{\text{PFM}}(\beta, \mathbf{z}) : q_{\text{PFM}}(\beta, \mathbf{z}) = q_{\text{PFM}}(\beta \mid \mathbf{z}) \prod_{i=1}^{n} q_{\text{PFM}}(z_i)\}$.

Motivation for the use of $\mathcal{Q}_{\text{PFM}}$: $q^*_{\text{MF}}(\beta, \mathbf{z}) = q^*_{\text{MF}}(\beta) \prod_{i=1}^{n} q^*_{\text{MF}}(z_i)$ belongs to $\mathcal{Q}_{\text{PFM}}$, and $p(\beta, \mathbf{z} \mid \mathbf{y}) = p(\beta \mid \mathbf{z})p(\mathbf{z} \mid \mathbf{y})$ with $p(\beta \mid \mathbf{z}) = \phi_p(\beta - \mathbf{V}\mathbf{X}^\intercal\mathbf{z}; \mathbf{V})$ and $p(\mathbf{z} \mid \mathbf{y}) \propto \phi_n(\mathbf{z}; \mathbf{I}_n + \nu_p^2 \mathbf{X}\mathbf{X}^\intercal) \prod_{i=1}^{n} \mathbb{1}[(2y_i - 1)z_i > 0]$ [Holmes and Held, 2006].

Proposition. Let $q^*_{\text{PFM}}(\beta, \mathbf{z})$ and $q^*_{\text{MF}}(\beta, \mathbf{z})$ be the optimal approximations for $p(\beta, \mathbf{z} \mid \mathbf{y})$, under PFM-VB and MF-VB, respectively. Then

$$\text{KL}[q^*_{\text{PFM}}(\beta, \mathbf{z}) \,||\, p(\beta, \mathbf{z} \mid \mathbf{y})] \leq \text{KL}[q^*_{\text{MF}}(\beta, \mathbf{z}) \,||\, p(\beta, \mathbf{z} \mid \mathbf{y})].$$

Main Theorem. The optimal joint approximating density $q^*_{\text{PFM}}(\beta, \mathbf{z})$ can be derived via a tractable CAVI relying on simple closed–form expressions and $q^*_{\text{PFM}}(\beta) = \int_{\mathbb{R}^n} q^*_{\text{PFM}}(\beta \mid \mathbf{z}) \prod_{i=1}^{n} q^*_{\text{PFM}}(z_i) d\mathbf{z} = \mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}[q^*_{\text{PFM}}(\beta \mid \mathbf{z})]$ of direct interest is the density of a SUN, which crucially relies on a diagonal $\mathbf{\Gamma} = \mathbf{I}_n$.

To be useful in practice, $q^*_{\text{PFM}}(\beta, \mathbf{z})$ should be simple to derive and the density $q^*_{\text{PFM}}(\beta) = \int_{\mathbb{R}^n} q^*_{\text{PFM}}(\beta|\mathbf{z}) \prod_{i=1}^n q^*_{\text{PFM}}(z_i)\text{d}\mathbf{z} = \mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}[q^*_{\text{PFM}}(\beta \mid \mathbf{z})]$ of direct interest should be available in tractable form.

To be useful in practice, $q^*_{\text{PFM}}(\beta, \mathbf{z})$ should be simple to derive and the density $q^*_{\text{PFM}}(\beta) = \int_{\mathbb{R}^n} q^*_{\text{PFM}}(\beta|\mathbf{z}) \prod_{i=1}^n q^*_{\text{PFM}}(z_i) d\mathbf{z} = \mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}[q^*_{\text{PFM}}(\beta \mid \mathbf{z})]$ of direct interest should be available in tractable form.

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

$$q^*_{\text{PFM}}(\beta \mid \mathbf{z}) = p(\beta \mid \mathbf{z}) = \phi_p(\beta - \mathbf{V}\mathbf{X}^\intercal \mathbf{z}; \mathbf{V}), \quad \mathbf{V} = (\nu_p^{-2}\mathbf{I}_p + \mathbf{X}^\intercal\mathbf{X})^{-1},$$

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

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

To be useful in practice, $q^*_{\text{PFM}}(\beta, \mathbf{z})$ should be simple to derive and the density $q^*_{\text{PFM}}(\beta) = \int_{\mathbb{R}^n} q^*_{\text{PFM}}(\beta|\mathbf{z}) \prod_{i=1}^n q^*_{\text{PFM}}(z_i)\mathrm{d}\mathbf{z} = \mathbb{E}_{q^*_{\text{PFM}}(\mathbf{z})}[q^*_{\text{PFM}}(\beta \mid \mathbf{z})]$ of direct interest should be available in tractable form.

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

$$q^*_{\text{PFM}}(\beta \mid \mathbf{z}) = p(\beta \mid \mathbf{z}) = \phi_p(\beta - \mathbf{V}\mathbf{X}^\intercal\mathbf{z}; \mathbf{V}), \quad \mathbf{V} = (\nu_p^{-2}\mathbf{I}_p + \mathbf{X}^\intercal\mathbf{X})^{-1},$$

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

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

The optimal parameters of the above densities can be obtained via a simple CAVI algorithm [at the same cost of MF–VB].

Approximation quality and computational efficiency

Fasano, **Durante**, Zanella (2022+), *Biometrika*

The factorized form for $q_{\text{PFM}}(\mathbf{z})$ leads to a SUN approximate density for β, with $\mathbf{\Gamma} = \mathbf{I}_n$. This allows tractable inference at an $\mathcal{O}(pn \cdot \min\{p, n\})$ cost.

Approximation quality and computational efficiency

Fasano, **Durante**, Zanella (2022+), *Biometrika*

The factorized form for $q_{\text{PFM}}(\mathbf{z})$ leads to a SUN approximate density for β, with $\boldsymbol{\Gamma} = \mathbf{I}_n$. This allows tractable inference at an $\mathcal{O}(pn \cdot \min\{p, n\})$ cost.

Theorem Under simple assumptions, $\text{KL}[q^*_{\text{PFM}}(\beta) \mid\mid p(\beta \mid \mathbf{y})] \xrightarrow{p} 0$ as $p \to \infty$
[quality of the approximation]

Approximation quality and computational efficiency

Fasano, **Durante**, Zanella (2022+), *Biometrika*

The factorized form for $q_{\text{PFM}}(\mathbf{z})$ leads to a SUN approximate density for β, with $\boldsymbol{\Gamma} = \mathbf{I}_n$. This allows tractable inference at an $\mathcal{O}(pn \cdot \min\{p, n\})$ cost.

Theorem Under simple assumptions, $\text{KL}[q^*_{\text{PFM}}(\beta) \mid\mid p(\beta \mid \mathbf{y})] \xrightarrow{p} 0$ as $p \to \infty$ [quality of the approximation]

Corollary. Let $\text{pr}(y_{\text{NEW}} = 1 \mid \mathbf{y}) = \int \Phi(\mathbf{x}_{\text{NEW}}^\intercal \beta) p(\beta|\mathbf{y}) \text{d}\beta$ be the exact posterior predictive probability for a new unit with predictors $\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p$. Then, under simple assumptions, $\sup_{\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p} |\text{pr}_{\text{PFM}}(y_{\text{NEW}} = 1 \mid \mathbf{y}) - \text{pr}(y_{\text{NEW}} = 1 \mid \mathbf{y})| \xrightarrow{p} 0$ as $p \to \infty$. Instead, $\liminf_{p \to \infty} \sup_{\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p} |\text{pr}_{\text{MF}}(y_{\text{NEW}} = 1|\mathbf{y}) - \text{pr}(y_{\text{NEW}} = 1|\mathbf{y})| > 0$ almost surely as $p \to \infty$ [quality of classification]

Approximation quality and computational efficiency

Fasano, **Durante**, Zanella (2022+), *Biometrika*

The factorized form for $q_{\text{PFM}}(\mathbf{z})$ leads to a SUN approximate density for β, with $\mathbf{\Gamma} = \mathbf{I}_n$. This allows tractable inference at an $\mathcal{O}(pn \cdot \min\{p, n\})$ cost.

Theorem Under simple assumptions, $\text{KL}[q^*_{\text{PFM}}(\beta) \,||\, p(\beta \mid \mathbf{y})] \xrightarrow{p} 0$ as $p \to \infty$ [quality of the approximation]

Corollary. Let $\text{pr}(y_{\text{NEW}} = 1 \mid \mathbf{y}) = \int \Phi(\mathbf{x}_{\text{NEW}}^{\intercal}\beta)p(\beta|\mathbf{y})\text{d}\beta$ be the exact posterior predictive probability for a new unit with predictors $\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p$. Then, under simple assumptions, $\sup_{\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p} |\text{pr}_{\text{PFM}}(y_{\text{NEW}} = 1 \mid \mathbf{y}) - \text{pr}(y_{\text{NEW}} = 1 \mid \mathbf{y})| \xrightarrow{p} 0$ as $p \to \infty$. Instead, $\liminf_{p \to \infty} \sup_{\mathbf{x}_{\text{NEW}} \in \mathbb{R}^p} |\text{pr}_{\text{MF}}(y_{\text{NEW}} = 1|\mathbf{y}) - \text{pr}(y_{\text{NEW}} = 1|\mathbf{y})| > 0$ almost surely as $p \to \infty$ [quality of classification]

Theorem. Let $q^{(t)}_{\text{PFM}}(\beta) = \int_{\mathbb{R}^n} q^{(t)}_{\text{PFM}}(\beta \mid \mathbf{z}) \prod_{i=1}^n q^{(t)}_{\text{PFM}}(z_i)\text{d}\mathbf{z}$ be the approximate density for β produced at iteration t by our CAVI. Then, under simple assumptions, $\text{KL}[q^{(1)}_{\text{PFM}}(\beta) \,||\, p(\beta \mid \mathbf{y})] \xrightarrow{p} 0$ as $p \to \infty$ [computational efficiency]

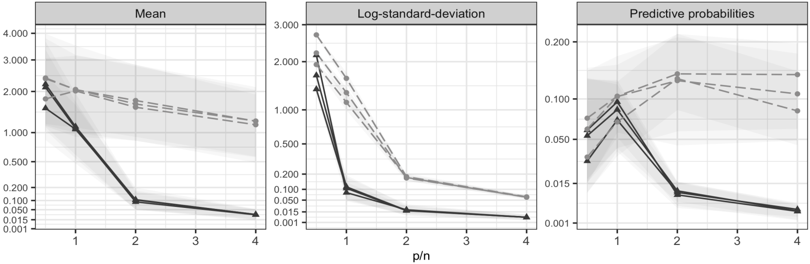
Simulation

We **evaluate accuracy in the approximation for three key functionals** of the posterior distribution for β, by comparing MF–VB and PFM–VB approximations for these quantities with the STAN estimates at varying (p, n) settings.

Simulation scenario: data $\mathbf{y}$ are simulated from probit regression with inputs x_{ij}, $[i = 1, \ldots, n,\ j = 1, \ldots, p]$ sampled from **independent standard normals** and coefficients β_j $[j = 1, \ldots, p]$ simulated from uniforms in the range $[-5, 5]$.

We **evaluate accuracy in the approximation for three key functionals** of the posterior distribution for β, by comparing MF–VB and PFM–VB approximations for these quantities with the STAN estimates at varying (p, n) settings.

Simulation scenario: data $\mathbf{y}$ are simulated from probit regression with inputs x_{ij}, $[i = 1, \ldots, n,\ j = 1, \ldots, p]$ sampled from **independent standard normals** and coefficients β_j $[j = 1, \ldots, p]$ simulated from uniforms in the range $[-5, 5]$.



Empirical evidence is in line with theory and shows that our asymptotic results are visible also in finite-dimensional $p > n$ settings.

Alzheimers' application

Large p, moderate n study on presence–absence of Alzheimer as a function of demographic data, genotype and assay results. In this application $n = 300$ and $p = 9036$ [we include interactions]. We consider $\beta \sim \mathrm{N}_{9036}(\mathbf{0}, 25 \cdot \mathbf{I}_{9036})$.

Alzheimers' application

Large p, moderate n study on presence–absence of Alzheimer as a function of demographic data, genotype and assay results. In this application $n = 300$ and $p = 9036$ [we include interactions]. We consider $\beta \sim \text{N}_{9036}(\mathbf{0}, 25 \cdot \mathbf{I}_{9036})$.

Computational performance. Runtimes required for posterior inference

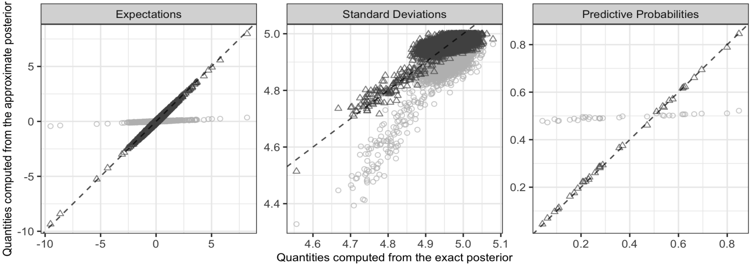
	STAN	EP	SUN	MF-VB	PFM-VB
Time [minutes]	> 360.00	> 360.00	92.27	0.04	0.04

Alzheimers' application

Large p, moderate n study on presence–absence of Alzheimer as a function of demographic data, genotype and assay results. In this application $n = 300$ and $p = 9036$ [we include interactions]. We consider $\beta \sim \mathrm{N}_{9036}(\mathbf{0}, 25 \cdot \mathbf{I}_{9036})$.

Computational performance. Runtimes required for posterior inference

	STAN	EP	SUN	MF-VB	PFM-VB
Time [minutes]	> 360.00	> 360.00	92.27	0.04	0.04



The models considered so far are special examples of a much **broader class of formulations** whose likelihood factorizes as

$$p(\mathbf{y} \mid \beta) = p(\mathbf{y}_1 \mid \beta)p(\mathbf{y}_0 \mid \beta) \propto \phi_{n_1}(\mathbf{y}_1 - \mathbf{X}_1\beta; \mathbf{\Sigma}_1)\Phi_{n_0}(\mathbf{y}_0 + \mathbf{X}_0\beta; \mathbf{\Sigma}_0). \quad (3)$$

Examples: probit, multivariate probit, multinomial probit, tobit, and others.

Further extensions

Anceschi, Fasano, **Durante**, Zanella (202–), *https://arxiv.org/abs/2206.08118*

The models considered so far are special examples of a much **broader class of formulations** whose likelihood factorizes as

$$p(\mathbf{y} \mid \beta) = p(\mathbf{y}_1 \mid \beta)p(\mathbf{y}_0 \mid \beta) \propto \phi_{n_1}(\mathbf{y}_1 - \mathbf{X}_1\beta; \mathbf{\Sigma}_1)\Phi_{n_0}(\mathbf{y}_0 + \mathbf{X}_0\beta; \mathbf{\Sigma}_0). \tag{3}$$

Examples: probit, multivariate probit, multinomial probit, tobit, and others.

Note: Recalling the results in the previous slides, the above likelihood is actually the kernel of a SUN density → **conjugacy** with SUN priors.

Anceschi, Fasano, **Durante**, Zanella (202–), *https://arxiv.org/abs/2206.08118*

The models considered so far are special examples of a much **broader class of formulations** whose likelihood factorizes as

$$p(\mathbf{y} \mid \beta) = p(\mathbf{y}_1 \mid \beta)p(\mathbf{y}_0 \mid \beta) \propto \phi_{n_1}(\mathbf{y}_1 - \mathbf{X}_1\beta; \mathbf{\Sigma}_1)\Phi_{n_0}(\mathbf{y}_0 + \mathbf{X}_0\beta; \mathbf{\Sigma}_0). \qquad (3)$$

Examples: probit, multivariate probit, multinomial probit, tobit, and others.

Note: Recalling the results in the previous slides, the above likelihood is actually the kernel of a SUN density → **conjugacy** with SUN priors.

Main Theorem. If $\beta \sim \text{SUN}_{p,q}(\boldsymbol{\xi}, \mathbf{\Omega}, \mathbf{\Delta}, \gamma, \mathbf{\Gamma})$ — meaning that the prior density of β is $p(\beta) \propto \phi_p(\beta - \xi; \mathbf{\Omega})\Phi_q(\gamma + \mathbf{\Delta}^\intercal\bar{\mathbf{\Omega}}^{-1}\omega^{-1}(\beta - \xi); \mathbf{\Gamma} - \mathbf{\Delta}^\intercal\bar{\mathbf{\Omega}}^{-1}\mathbf{\Delta})$ — and $p(\mathbf{y} \mid \beta)$ has likelihood (3), then

$$(\beta \mid \mathbf{y}) \sim \text{SUN}_{p,q+n_0}(\boldsymbol{\xi}_{\text{POST}}, \mathbf{\Omega}_{\text{POST}}, \mathbf{\Delta}_{\text{POST}}, \gamma_{\text{POST}}, \mathbf{\Gamma}_{\text{POST}}),$$

where $\boldsymbol{\xi}_{\text{POST}}, \mathbf{\Omega}_{\text{POST}}, \mathbf{\Delta}_{\text{POST}}, \gamma_{\text{POST}}$, and $\mathbf{\Gamma}_{\text{POST}}$ are simple analytical functions of $\xi, \mathbf{\Omega}, \mathbf{\Delta}, \gamma, \mathbf{\Gamma}$ and $\mathbf{y}_1, \mathbf{X}_1, \mathbf{\Sigma}_1, \mathbf{y}_0, \mathbf{X}_0, \mathbf{\Sigma}_0$.

Further extensions

Anceschi, Fasano, **Durante**, Zanella (202–), *https://arxiv.org/abs/2206.08118*

The models considered so far are special examples of a much **broader class of formulations** whose likelihood factorizes as

$$p(\mathbf{y} \mid \beta) = p(\mathbf{y}_1 \mid \beta) p(\mathbf{y}_0 \mid \beta) \propto \phi_{n_1}(\mathbf{y}_1 - \mathbf{X}_1\beta; \boldsymbol{\Sigma}_1)\Phi_{n_0}(\mathbf{y}_0 + \mathbf{X}_0\beta; \boldsymbol{\Sigma}_0). \quad (3)$$

Examples: probit, multivariate probit, multinomial probit, tobit, and others.

Note: Recalling the results in the previous slides, the above likelihood is actually the kernel of a SUN density → **conjugacy** with SUN priors.

Main Theorem. If $\beta \sim \text{SUN}_{p,q}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \gamma, \boldsymbol{\Gamma})$ — meaning that the prior density of β is $p(\beta) \propto \phi_p(\beta - \xi; \boldsymbol{\Omega})\Phi_q(\gamma + \boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\omega^{-1}(\beta - \xi); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta})$ — and $p(\mathbf{y} \mid \beta)$ has likelihood (3), then

$$(\beta \mid \mathbf{y}) \sim \text{SUN}_{p,q+n_0}(\boldsymbol{\xi}_{\text{POST}}, \boldsymbol{\Omega}_{\text{POST}}, \boldsymbol{\Delta}_{\text{POST}}, \gamma_{\text{POST}}, \boldsymbol{\Gamma}_{\text{POST}}),$$

where $\boldsymbol{\xi}_{\text{POST}}, \boldsymbol{\Omega}_{\text{POST}}, \boldsymbol{\Delta}_{\text{POST}}, \gamma_{\text{POST}}$, and $\boldsymbol{\Gamma}_{\text{POST}}$ are simple analytical functions of $\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \gamma, \boldsymbol{\Gamma}$ and $\mathbf{y}_1, \mathbf{X}_1, \boldsymbol{\Sigma}_1, \mathbf{y}_0, \mathbf{X}_0, \boldsymbol{\Sigma}_0$.

Consequence: All computational and inference methods previously developed can be applied to a broad class of routinely–implemented models.

Other interesting results

Cao, **Durante**, Genton (2022+), *Journal of Computational and Graphical Statistics*

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 00, NO. 0, 1–12
https://doi.org/10.1080/10618600.2022.2036614

Taylor & Francis
Taylor & Francis Group

Check for updates

Scalable Computation of Predictive Probabilities in Probit Models with Gaussian Process Priors

Jian Cao[a], Daniele Durante[b], and Marc G. Genton[a]

[a]Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia; [b]Department of Decision Sciences and Bocconi Institute for Data Science and Analytics, Bocconi University, Milano, Italy

ABSTRACT

Predictive models for binary data are fundamental in various fields, and the growing complexity of modern applications has motivated several flexible specifications for modeling the relationship between the observed predictors and the binary responses. A widely-implemented solution is to express the probability parameter via a probit mapping of a Gaussian process indexed by predictors. However, unlike for continuous settings, there is a lack of closed-form results for predictive distributions in binary models with Gaussian process priors. Markov chain Monte Carlo methods and approximation strategies provide common solutions to this problem, but state-of-the-art algorithms are either computationally intractable or inaccurate in moderate-to-high dimensions. In this article, we aim to cover this gap by deriving closed-form expressions for the predictive probabilities in probit Gaussian processes that rely either on cumulative distribution functions of multivariate Gaussians or on functionals of multivariate truncated normals. To evaluate these quantities we develop novel scalable solutions based on tile-low-rank Monte Carlo methods for computing multivariate Gaussian probabilities, and on mean-field variational approximations of multivariate truncated normals. Closed-form expressions for the marginal likelihood and for the posterior distribution of the Gaussian process are also discussed. As shown in simulated and real-world empirical studies, the proposed methods scale to dimensions where state-of-the-art solutions are impractical.

ARTICLE HISTORY
Received September 2020
Revised November 2021

KEYWORDS
Binary data; Gaussian process; Multivariate truncated normal; Probit model; Unified skew-normal; Variational Bayes

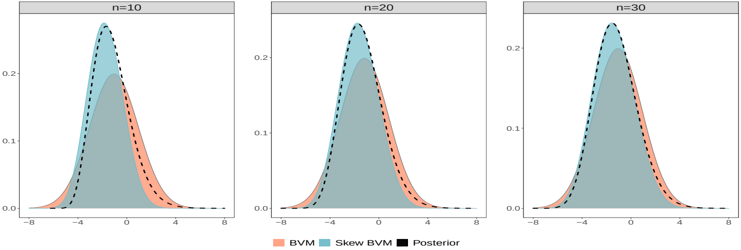
Main result: Derive closed–form expressions for the predictive probabilities in probit Gaussian processes that rely on ratios of cdfs of multivariate Gaussians and develop new scalable solutions based on **tile–low–rank Monte Carlo** methods and **separation–of–variables estimator** [Genz, 1992] for computing ratios of Gaussian cdfs with theoretical accuracy guarantees

Bernstein–Von Mises theorem

Bernstein–Von Mises theorem [in short]: under regularity conditions, the **total variation distance** between the posterior distribution and a suitably–defined Gaussian distribution converges to 0 in probability.

Bernstein–Von Mises theorem

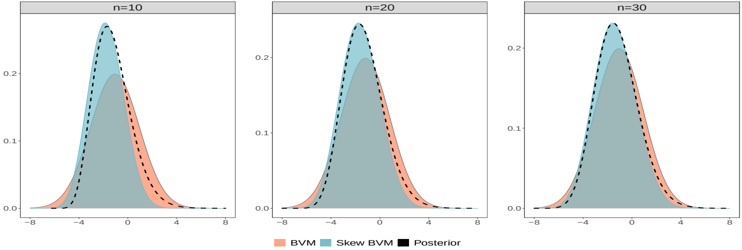
Bernstein–Von Mises theorem [in short]: under regularity conditions, the **total variation distance** between the posterior distribution and a suitably–defined Gaussian distribution converges to 0 in probability.



However: This limiting behavior may require a large sample size before becoming visible. In fact, the posterior distribution is often skewed in practice.

Bernstein–Von Mises theorem

Bernstein–Von Mises theorem [in short]: under regularity conditions, the **total variation distance** between the posterior distribution and a suitably–defined Gaussian distribution converges to 0 in probability.



However: This limiting behavior may require a large sample size before becoming visible. In fact, the posterior distribution is often skewed in practice.
Conjecture: Adopting as limiting law a skewed generalization of the Gaussian distribution, we might obtain substantially more accurate/stronger results.

Skewed Bernstein–Von Mises theorem

Pozza, **Durante**, Szabo (2022+), *soon online*

Let $\{\mathbf{y}_i\}_{i=1}^n$ be a sequence of independent random variables with probability measure $P_{\boldsymbol{\theta}_0}^{(n)} \in \{P_{\boldsymbol{\theta}}^{(n)}, \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$. Moreover, let $\ell(\boldsymbol{\theta})$ be the log–likelihood and $\ell^{(1)} = [\ell_r^{(1)}]$, $\ell^{(2)} = [\ell_{rs}^{(2)}]$, $\ell^{(3)} = [\ell_{rst}^{(3)}]$ its first three derivatives at $\boldsymbol{\theta}_0$.

Skewed Bernstein–Von Mises theorem

Pozza, **Durante**, Szabo (2022+), *soon online*

Let $\{\mathbf{y}_i\}_{i=1}^n$ be a sequence of independent random variables with probability measure $P_{\theta_0}^{(n)} \in \{P_{\theta}^{(n)}, \theta \in \Theta \subseteq \mathbb{R}^p\}$. Moreover, let $\ell(\theta)$ be the log–likelihood and $\ell^{(1)} = [\ell_r^{(1)}]$, $\ell^{(2)} = [\ell_{rs}^{(2)}]$, $\ell^{(3)} = [\ell_{rst}^{(3)}]$ its first three derivatives at θ_0.

Theorem: Under regularity conditions on the log–likelihood ratio and its derivatives, if the map $\theta \to P_{\theta}^{(n)}$ is one–to–one, θ_0 is an inner point of Θ and the prior measure $P(\theta)$ is absolutely continuous with bounded and positive density in a neighborhood of θ_0, then

$$|| \, P(\cdot \mid \mathbf{y}^{(n)}) - P_{se}(\cdot) \, ||_{\text{TV}} = O_p(\{\log n\}^{p/2+3}/n)$$

where $P_{se}(\mathbb{A}) = \int_{\mathbb{A}} p_{se}(\bar{\theta})d\bar{\theta}$ for $\mathbb{A} \subset \mathbb{R}^p$, $\bar{\theta} = \sqrt{n}(\theta - \theta_0)$ and $p_{se}(\bar{\theta})$ is the density of a suitably–defined skew–symmetric distribution [Azzalini & Regoli, 2012]. Specifically, $p_{se}(\bar{\theta}) = 2\phi_p(\bar{\theta}; \xi_n, \Omega_n)\Phi\{\alpha_n(\bar{\theta})\}$, where $\alpha_n(\cdot)$ is an odd function.

Skewed Bernstein–Von Mises theorem

Pozza, **Durante**, Szabo (2022+), *soon online*

Let $\{\mathbf{y}_i\}_{i=1}^n$ be a sequence of independent random variables with probability measure $P_{\theta_0}^{(n)} \in \{P_{\theta}^{(n)}, \theta \in \Theta \subseteq \mathbb{R}^p\}$. Moreover, let $\ell(\theta)$ be the log–likelihood and $\ell^{(1)} = [\ell_r^{(1)}]$, $\ell^{(2)} = [\ell_{rs}^{(2)}]$, $\ell^{(3)} = [\ell_{rst}^{(3)}]$ its first three derivatives at θ_0.

Theorem: Under regularity conditions on the log–likelihood ratio and its derivatives, if the map $\theta \to P_{\theta}^{(n)}$ is one–to–one, θ_0 is an inner point of Θ and the prior measure $P(\theta)$ is absolutely continuous with bounded and positive density in a neighborhood of θ_0, then

$$|| \, P(\cdot \mid \mathbf{y}^{(n)}) - P_{se}(\cdot) \, ||_{\text{TV}} = O_p(\{\log n\}^{p/2+3}/n)$$

where $P_{se}(\mathbb{A}) = \int_{\mathbb{A}} p_{se}(\bar{\theta}) d\bar{\theta}$ for $\mathbb{A} \subset \mathbb{R}^p$, $\bar{\theta} = \sqrt{n}(\theta - \theta_0)$ and $p_{se}(\bar{\theta})$ is the density of a suitably–defined skew–symmetric distribution [Azzalini & Regoli, 2012]. Specifically, $p_{se}(\bar{\theta}) = 2\phi_p(\bar{\theta}; \xi_n, \Omega_n)\Phi\{\alpha_n(\bar{\theta})\}$, where $\alpha_n(\cdot)$ is an odd function.

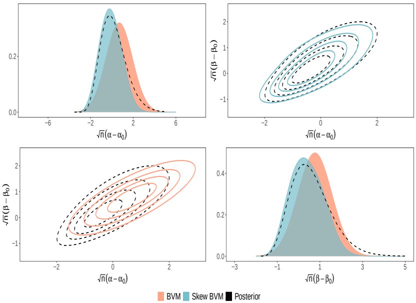
Remark: In the above theorem, the quantities ξ_n, Ω_n and $\alpha_n(\cdot)$ are simple analytical functions of $\ell^{(1)} = [\ell_r^{(1)}]$, $\ell^{(2)} = [\ell_{rs}^{(2)}]$, $\ell^{(3)} = [\ell_{rst}^{(3)}]$ and the prior.

Skew–modal approximation

Pozza, **Durante**, Szabo (2022+), *soon online*

Skew–modal approximation

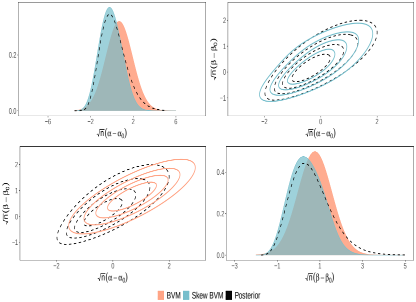
Pozza, **Durante**, Szabo (2022+), *soon online*



Simulation with $n = 15$, $y_i \overset{iid}{\sim} \text{Ga}(\alpha, \beta)$, $\alpha \sim \text{Ga}(2)$ and $\beta \sim \text{Ga}(2)$.

Skew–modal approximation

Pozza, **Durante**, Szabo (2022+), *soon online*



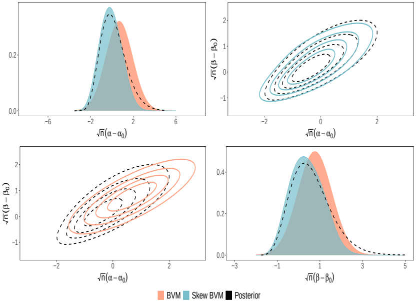
Simulation with $n = 15$, $y_i \overset{iid}{\sim} \text{Ga}(\alpha, \beta)$, $\alpha \sim \text{Ga}(2)$ and $\beta \sim \text{Ga}(2)$.

Comment. We improve the approximation accuracy relative to classical BvM. However, both approximations require $\boldsymbol{\theta}_0$, which is not available in practice.

Solution. Modal approximation based on a **skew–symmetric density** rather than a Gaussian one [recall Laplace approximation]

Skew–modal approximation

Pozza, **Durante**, Szabo (2022+), *soon online*

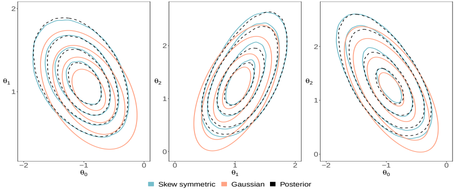


Simulation with $n = 15$, $y_i \overset{iid}{\sim} \text{Ga}(\alpha, \beta)$, $\alpha \sim \text{Ga}(2)$ and $\beta \sim \text{Ga}(2)$.

Comment. We improve the approximation accuracy relative to classical BvM. However, both approximations require $\boldsymbol{\theta}_0$, which is not available in practice.

Solution. Modal approximation based on a **skew–symmetric density** rather than a Gaussian one [recall Laplace approximation]

Skew–modal approximation [provably more accurate than Laplace]: Let $\tilde{\ell}$ denote the log–posterior at its MAP $\tilde{\boldsymbol{\theta}}$, then we approximate $p(\boldsymbol{\theta} \mid \mathbf{y}^{(n)})$ via $2\phi_p(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Omega}})\Phi\{\tilde{\alpha}(\boldsymbol{\theta})\}$ where $\tilde{\boldsymbol{\Omega}}$ and $\tilde{\alpha}(\cdot)$ are simple functions of $\tilde{\ell}^{(2)}$, $\tilde{\ell}^{(3)}$ and $\tilde{\boldsymbol{\theta}}$.



Conclusion

Main message: Skew–normals and related families [**Azzalini** & co–authors] play a key role in Bayesian inference, which has been partially overlooked to date [Exception: **Liseo** & co–authors]. The advancements presented open new avenues for improved posterior inference via novel closed–form expressions, new Monte Carlo methods, and more accurate and scalable approximations.

Conclusion

Main message: Skew–normals and related families [**Azzalini** & co–authors] play a key role in Bayesian inference, which has been partially overlooked to date [Exception: **Liseo** & co–authors]. The advancements presented open new avenues for improved posterior inference via novel closed–form expressions, new Monte Carlo methods, and more accurate and scalable approximations.

The above results also motivate **further extensions**.

- Further improve the skew–modal approximation in terms of accuracy
- Explore conjugacy in broader classes [of models and skewed prior]
- Explore more complex models building on such representations; i.e. BART

Conclusion

Main message: Skew–normals and related families [**Azzalini** & co–authors] play a key role in Bayesian inference, which has been partially overlooked to date [Exception: **Liseo** & co–authors]. The advancements presented open new avenues for improved posterior inference via novel closed–form expressions, new Monte Carlo methods, and more accurate and scalable approximations.

The above results also motivate **further extensions**.

- Further improve the skew–modal approximation in terms of accuracy
- Explore conjugacy in broader classes [of models and skewed prior]
- Explore more complex models building on such representations; i.e. BART

Thank you for the attention!

https://danieledurante.github.io/web/

https://github.com/danieledurante

I would like to acknowledge the support from MIUR–PRIN 2017 project SELECT [20177BRJXS] in the preparation of some of these works.