Skew 2026 Workshop, Padova 7-9 January 2026

Beyond SUN distributions for Bayesian binary regression

Brunero Liseo

MEMOTEF Department - Sapienza University of Rome
paolo.onorati@dauphine.psl.eu

Joint work with Paolo Onorati

CEREMADE - Université Paris Dauphine-PSL

Some memories

- I started working on the scalar Skew Normal in its simplest form

$$f(z, \lambda) = 2\varphi(z)\Phi(\lambda z), \qquad \lambda > 0; z \in \mathbb{R}$$

  in 1987 for my Laurea's dissertation
- I was not able to notice, in a formal way, the *strange* behaviour of the likelihood function
- For a while, I was suspecting that my code was wrong (two pages of code, which, today, would correspond to
  `pnorm(lambda * z)`

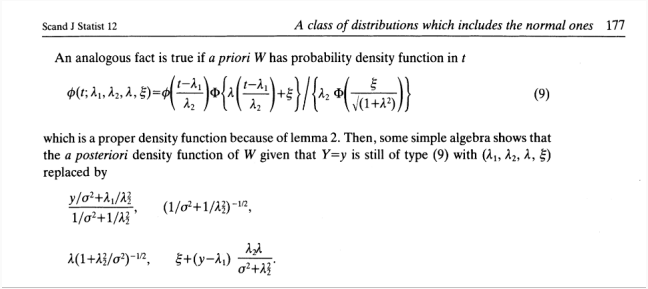
Over the years . . .

- I kept trying to find good inferential strategies for estimating the parameters of some skew-symmetric statistical model

[1] Azzalini, A. (1985). A class of distributions which includes the normal ones. Scandinavian journal of statistics, 171-178.

Over the years . . .

- I kept trying to find good inferential strategies for estimating the parameters of some skew-symmetric statistical model
- I ignored the other result in Adelchi's paper[1]



Scand J Statist 12 *A class of distributions which includes the normal ones* 177

An analogous fact is true if *a priori* W has probability density function in t

$$\phi(t;\lambda_1,\lambda_2,\lambda,\xi)=\phi\left(\frac{t-\lambda_1}{\lambda_2}\right)\Phi\left\{\lambda\left(\frac{t-\lambda_1}{\lambda_2}\right)+\xi\right\}\Big/\left\{\lambda_2\,\Phi\left(\frac{\xi}{\sqrt{(1+\lambda^2)}}\right)\right\} \qquad (9)$$

which is a proper density function because of lemma 2. Then, some simple algebra shows that the *a posteriori* density function of W given that $Y=y$ is still of type (9) with $(\lambda_1, \lambda_2, \lambda, \xi)$ replaced by

$$\frac{y/\sigma^2+\lambda_1/\lambda_2^2}{1/\sigma^2+1/\lambda_2^2}, \qquad (1/\sigma^2+1/\lambda_2^2)^{-1/2},$$

$$\lambda(1+\lambda_2^2/\sigma^2)^{-1/2}, \qquad \xi+(y-\lambda_1)\,\frac{\lambda_2\lambda}{\sigma^2+\lambda_2^2}.$$

[1] Azzalini, A. (1985). A class of distributions which includes the normal ones. Scandinavian journal of statistics, 171-178.

SUN class as a conjugate prior

- Durante (2019) discovered a central role of the SUN density in Bayesian probit models.
- The result has been generalised in many directions by Daniele and his collaborators

Starting from a normal prior for the coefficients $\beta \sim N_p(\xi, \Omega)$, the posterior for β after producing a probit likelihood, belongs to the SUN family

$$\beta | y, X \sim SUN_{p,n}(\xi^*, \Omega^*, \Delta^*, \tau^*, \Gamma^*)$$

Our Starting Point

- Is it possible to have a manageable posterior in the `logit` and (maybe ...) other cases? (Onorati and Liseo, 2025)
- Can we go beyond a parametric approach? (work in progress)

Outline

Ingredients:

- The Unified Skew Normal (SUN) class of densities
- Scale mixtures of Gaussian random variables
- Kolmogorov distribution

Prequel

The Unified Skew-Normal density has been introduced by Arellano-Valle and Azzalini (2006).

It has several representations: we see it as a multivariate Gaussian with a latent selection mechanism. We say β is a d-dimensional SUN random vector, i.e. $\beta \sim \mathrm{SUN}_{p,m}(\xi, \Omega, \Delta, \tau, \Gamma)$ if

$$\beta = \xi + \mathrm{diag}^{1/2}(\Omega) Z \mid U + \tau > 0_m$$

with

$$\begin{bmatrix} Z \\ U \end{bmatrix} \sim N_{p+m}\left(\begin{bmatrix} 0_p \\ 0_m \end{bmatrix}, \begin{bmatrix} \bar{\Omega} & \Delta \\ \Delta' & \Gamma \end{bmatrix} \right),$$

$\xi \in \mathbb{R}^p, \tau \in \mathbb{R}^m, \Gamma$ is a m-correlation matrix, Ω is a p-covariance matrix, Δ is $p \times m$ matrix and $\bar{\Omega} = \mathrm{diag}^{-\frac{1}{2}}(\Omega)\, \Omega\, \mathrm{diag}^{-\frac{1}{2}}(\Omega)$.

When $m = 1$ and $\tau = 0$, it is the standard Skew-Normal density.

SUN Family and Probit Model

Durante (2019) discovered a central role of the SUN density in Bayesian probit models.

Starting from a SUN prior for the coefficients $\beta \sim \mathrm{SUN}_{p,m}(\xi, \Omega, \Delta, \tau, \Gamma)$ in a probit model, the resulting posterior still belongs to the SUN family.

$$\beta | Y = y \sim \mathrm{SUN}_{p,m+n}(\xi, \Omega, \Delta^\star, \tau^\star, \Gamma^\star)$$

Remarks:

- The previous stochastic representation can be suitably used for exact posterior sampling
- The algorithm is particularly efficient if $m + n \leq 100$ or $p >> m + n$ (Botev, 2017)
- A normal prior is a special case of the SUN distribution. Then, starting from a Gaussian prior, the resulting posterior is SUN

A Different Representation

We start from a different representation of a SUN distribution

$$\beta = \xi + \mathrm{diag}^{\frac{1}{2}}(\Omega) Z \mid T \leq AZ + b\,, \tag{1}$$

with $A \in \mathbb{R}^{m \times p}$, $b \in \mathbb{R}^m$.

This way, $T \perp\!\!\!\perp Z$ and $T \sim N_m(0, \Theta)$, with

$$\begin{aligned}
\Theta &= \mathrm{diag}^{-\frac{1}{2}} \left(\Gamma - \Delta' \bar{\Omega}^{-1} \Delta\right) \left(\Gamma - \Delta' \bar{\Omega}^{-1} \Delta\right) \mathrm{diag}^{-\frac{1}{2}} \left(\Gamma - \Delta' \bar{\Omega}^{-1} \Delta\right)\,, \\
A &= \mathrm{diag}^{-\frac{1}{2}} \left(\Gamma - \Delta' \bar{\Omega}^{-1} \Delta\right) \Delta' \bar{\Omega}^{-1}\,, \\
b &= \mathrm{diag}^{-\frac{1}{2}} \left(\Gamma - \Delta' \bar{\Omega}^{-1} \Delta\right) \tau\,,
\end{aligned}$$

and we denote $\beta \sim \mathrm{SUN}^{*}_{p,m}(\Theta, A, b, \xi, \Omega)$.

Extending the SUN Family

- We construct a larger class of densities, the perturbed SUN (pSUN) by replacing φ and Φ with scale mixtures of Gaussian densities.
- This is done to find a more general conjugacy in the Bayesian analysis of binary regression models.

Assume that

$$Z|W \sim N_p(0_p, \bar{\Omega}_W)\,,\; T|V \sim N_m(0_m, \Theta_V)\,,$$

with

$$\bar{\Omega}_W = \mathrm{diag}^{\frac{1}{2}}(W)\,\bar{\Omega}\,\mathrm{diag}^{\frac{1}{2}}(W)\;,\;\; \Theta_V = \mathrm{diag}^{\frac{1}{2}}(V)\,\Theta\,\mathrm{diag}^{\frac{1}{2}}(V)$$

$$V \sim Q_V(\cdot) \quad \perp\!\!\!\perp \quad W \sim Q_W(\cdot)$$

The pSUN class is defined as the expression (1) with the above assumptions on Z and T. Then,

$$\mathrm{pSUN}_{p,m}\,(Q_V, \Theta, A, b, Q_W, \xi, \Omega)\;.$$

The Density of a pSUN

Let $\beta \sim \mathrm{pSUN}_{p,m}(Q_V, \Theta, A, b, Q_W, \Omega, \xi)$. Then

$$f_\beta(\beta) = \varphi_{\Omega,Q_W}(\beta - \xi) \frac{\Phi_{\Theta,Q_V}\left(A\,\mathrm{diag}^{-\frac{1}{2}}(\Omega)(\beta - \xi) + b\right)}{\Psi_{Q_V,\Theta,A,Q_W,\bar{\Omega}}(b)}, \tag{2}$$

with

$$\varphi_{\Sigma,Q}(u) = \int_{\mathbb{R}^p} \prod_{i=1}^{p} \left(W_i^{-\frac{1}{2}}\right) \phi_\Sigma\left(\mathrm{diag}^{-\frac{1}{2}}(W)\,u\right) dQ(W),$$

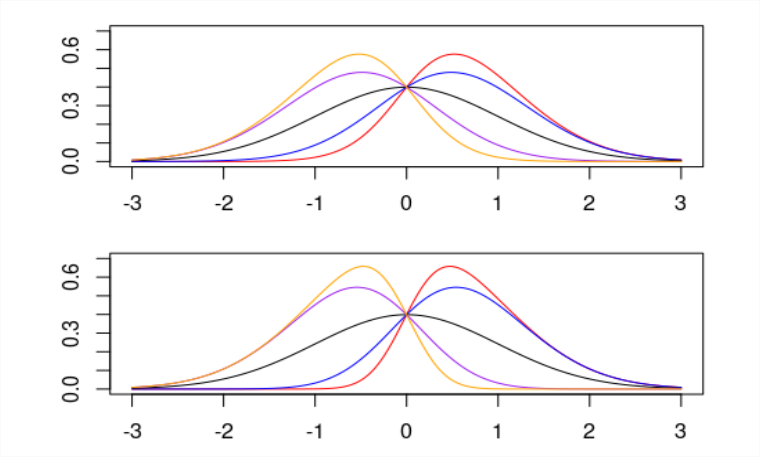
$$\Phi_{\Sigma,Q}(u) = \int_{\mathbb{R}^p} \Phi_\Sigma\left(\mathrm{diag}^{-\frac{1}{2}}(W)\,u\right) dQ(W),$$

and

$$\Psi_{Q_V,\Theta,A,Q_W,\bar{\Omega}}(b) = \mathrm{P}(T - AZ \leq b),$$
$$T \sim \Phi_{\Theta,Q_V}(\cdot) \perp\!\!\!\perp Z \sim \Phi_{\bar{\Omega},Q_W}(\cdot).$$

Some pSUN densities

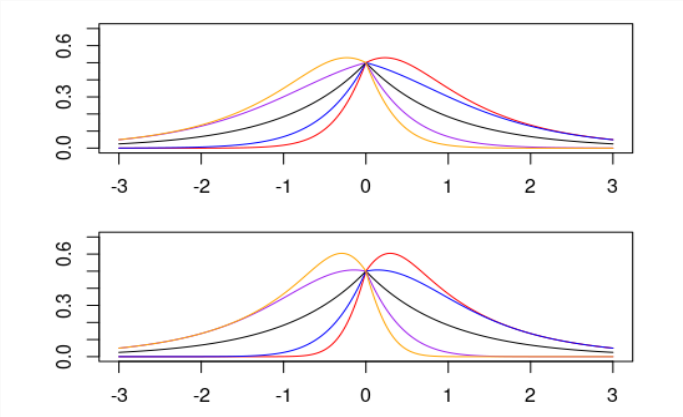
Logit: top: N(0,1); $V \sim LK(\cdot); W = 1, A = 3, b = 0$ $V \sim LK(\cdot); W = 1, A = 1.5, b = 0$



Probit: bottom: N(0,1); $V = W = 1, A = 3, b = 0$ $V = W = 1, A = 1.5, b = 0$

Some pSUN densities

top: Lapl(0,1) ; $V \sim LK(\cdot); W \sim Exp(0.5), A = 3, b = 0;$ $V \sim LK(\cdot); W \sim Exp(0.5), A = 1.5, b = 0$



bottom: Lapl(0,1) $V = 1, W \sim Exp(0.5), A = 3, b = 0;$ $V = 1, W \sim Exp(0.5), A = 1.5, b = 0$

Linear Symmetric Binary Regression Model

Consider a general version of the model as

$$Y_i|p_i \overset{\text{ind}}{\sim} Be(p_i)\,,\ \forall\, i = 1, 2, \ldots, n\,; \quad p_i = \Lambda\big(\eta(X_i)\big)\,,$$

- $\Lambda : \mathbb{R} \to [0, 1]$ is the link function,
- $\eta(\cdot)$ is a calibration function,
- $X_i \in \mathbb{R}^p$ is the i-th row of the design matrix X.

Typically, $\Lambda(\cdot)$ is a scalar CDF, symmetric about 0, and $\eta(x)$ takes the simple linear form, $x'\beta$; call it a linear symmetric binary regression model (LSBR).

Set $\Lambda_n(y) = \prod_{i=1}^n \Lambda(y_i)$ and $B_y = [2\,\mathrm{diag}(y) - I_n]$ for $y \in \{0, 1\}^n$, the likelihood function of a LSBR is

$$L\,(\beta; y) = \Lambda_n(B_y X\beta)\,.$$

Conjugacy for Linear Symmetric Binary Regression (LSBR)

Proposition 1

Consider a Bayesian LSBR model and assume

$$\beta \sim \mathrm{pSUN}_{p,m}(Q_V, \Theta, A, b, Q_W, \xi, \Omega).$$

If the link function is of the form $\Lambda(x) = \int_0^\infty \Phi\left(\frac{x}{\sqrt{v}}\right) dQ_{V_0}(v)$,

$$\beta | Y = y \sim \mathrm{pSUN}_{p,m+n}\left(Q_V^\star, \Theta^\star, A^\star, b^\star, Q_W, \xi, \Omega\right),$$

with

$$\Theta^\star = \begin{bmatrix} \Theta & 0_{m\times n} \\ 0_{n\times m} & I_n \end{bmatrix}; A^\star = \begin{bmatrix} A & 0_{m\times p} \\ 0_{n\times p} & B_y X \,\mathrm{diag}^{-\frac{1}{2}}(\Omega) \end{bmatrix}; b^\star = \begin{bmatrix} b \\ B_y X \xi \end{bmatrix},$$

and $Q_V^\star\left([x_1, x_2]'\right) = Q_V(x_1) \prod_{i=1}^n Q_{V_0}(x_{2,i})$

Bayesian Logistic Regression

- Logistic regression is a special case of LSBR
- The logistic CDF admits a representation in terms of a scale mixture of Gaussian distributions; see Andrews and Mallows (1974) and Stefanski (1991).

In fact,

$$T_i | \widetilde{V}_i \sim N(0, 4\widetilde{V}_i^2) \text{ and } \widetilde{V}_i \sim \text{Kolmogorov} \Longrightarrow T_i \sim \textit{Logis}(0,1)$$

that is

$$f_{T_i}(t) = \frac{\exp(-t)}{\left(1+\exp(-t)\right)^2}, \; t \in \mathbb{R}.$$

Kolmogorov's Distribution

We use the logistic Kolmogorov distribution:

$$V_i = 4\widetilde{V}_i^2 \ , \ \widetilde{V}_i \sim \text{Kolmogorov}$$

denoted by $V_i \sim \text{LK}(\cdot)$; the density is $f_{V_i}(v) = q_{Kol}(\sqrt{v}/2)/(4\sqrt{v})$, and

$$q_{Kol}(x) = \begin{cases} \dfrac{\sqrt{2\pi}}{x^2} \displaystyle\sum_{k=1}^{+\infty} \left(\frac{(2k-1)^2\pi^2}{4x^2} - 1 \right) \exp\left(-\frac{(2k-1)^2\pi^2}{8x^2} \right) , 0 < x \leq x_0 \\ 8x \displaystyle\sum_{k=1}^{+\infty} (-1)^{k-1} k^2 \exp\left(-2k^2x^2 \right) , \qquad x \geq x_0 \end{cases},$$

We set $x_0 = 1.207$ and truncate the series at 15-th term. This way the error is always less than 2.23×10^{-308}; see Onorati and Liseo (2022) for details.

Computation

In order to perform posterior computation, we use two different strategies provided by the stochastic representation of a pSUN distribution:

- Gibbs sampler when the prior is a scale mixture of Gaussian densities
  (Skip this part)
- importance sampling when the prior is Gaussian

Gibbs Sampler: Some Remarks (I)

- Holmes and Held (2006) have already used a very similar representation within a data-augmentation Gibbs algorithm for several models including logistic regression.
- Our approach and the one in Holmes and Held (2006) share some characteristics in the binary logistic case although we introduced improvements in terms of speed and mixing.
- We do: $V, W|T, Z$ and then $T, Z|V, W$
  Holmes and Held (2006) did: $V, W|T, Z$; then $T - AZ|Z, V, W$ and then $Z|T - AZ, V, W$
  in both cases, $\beta = \xi + \text{diag}^{1/2}(\Omega)Z$.

Gibbs Sampler: Some Remarks (II)

- One must be able to sample from the full conditional distributions of V and W.
- W: this is relatively simple when $p_\beta(\cdot)$ either has an elliptical structure or it has independent components. For example, the symmetric GH Barndorff-Nielsen (1977) class of priors satisfies the elliptical constraint and corresponds to $m = 0$. Instead, $m = 1 \Longrightarrow$ new skew version of the GH family.
- V: the first m component depend on the prior.

Gibbs Sampler: Some Remarks (III)

The hard step is "how to sample" from $V_i|T_i, i = m+1, m+2, \ldots, m+n$ because they depend on the link function. Focus on the last n components of $V|T$:

- They are always independent on the first m components of $V|T$ and they are also mutually independent.
- They depend on the specific link, and in the logistic case, they are the posterior density of a logistic Kolmogorov.
- We adopt an acceptance-rejection algorithm, with a *calibrated* inverse gamma proposal.

Gibbs Sampler: Some Remarks (IV)

Here

- $\widetilde{V} \sim \text{Inv.Gamma}(\alpha, \pi^2/2)$ and $\widetilde{T}|\widetilde{V} = \widetilde{v} \sim N(0, \widetilde{v})$.
- $\widetilde{V}|\widetilde{T} = t \sim \text{Inv.Gamma}\Big(\alpha + 1/2, (\pi^2 + t^2)/2\Big)$ (the proposal density).
- $f_{\widetilde{T}}(t) = \dfrac{\Gamma(\frac{2\alpha+1}{2})}{\Gamma(\alpha)\sqrt{\pi^3}} \left(1 + \dfrac{t^2}{\pi^2}\right)^{-\frac{2\alpha+1}{2}}$ (Student-t density).

Gibbs Sampler: Some Remarks (V)

Proposition 2

Let $V_i \sim \mathrm{LK}(\cdot)$, $T_i|V_i \sim N(0, V_i)$, $\widetilde{V} \sim \mathrm{Inv.Gamma}(\alpha, \pi^2/2)$ *and* $\widetilde{T}|\widetilde{V} \sim N(0, \widetilde{V})$*; set* $\alpha > 3/2$*. Then the ratio* $f_{V_i}(v|T_i = t)/f_{\widetilde{V}}(v|\widetilde{T} = t)$ *is bounded above by*

$$M^* \frac{f_{\widetilde{T}}(t)}{f_{T_i}(t)}, \tag{3}$$

with $M^* = \max\left(\max_{0<v\leq v^*} \delta_1(v), \max_{v>v^*} \delta_2(v)\right)$, $v^\star \in (1/2, 18\pi^2/11)$,

$$\begin{bmatrix}\delta_1(v)\\ \delta_2(v)\end{bmatrix} = \begin{cases} \dfrac{\sqrt{2\pi^5}\Gamma(\alpha)}{(\pi^2/2)^\alpha} v^{\alpha-\frac{3}{2}} & \text{if} \quad 0 < v \leq v^* \\ \dfrac{\Gamma(\alpha)}{(\pi^2/2)^\alpha} v^{\alpha+1} \exp\left(\dfrac{\pi^2}{2v} - \dfrac{v}{2}\right) & \text{if} \quad v > v^* \end{cases},$$

$$\arg\max_{0<v\leq v^*} \delta_1(v) = v^*,$$

$$\arg\max_{v>v^*} \delta_2(v) = \begin{cases} 1+\alpha+\sqrt{(1+\alpha)^2 - \pi^2} & \text{if} \quad \alpha \geq \pi - 1 \\ v^* & \text{otherwise} \end{cases}.$$

Gibbs Sampler: Some Remarks (VI)

Indeed, it is possible to show that the MGF and expected value of the posterior logistic Kolmogorov are respectively

$$M_{V_i}(u|T_i = t) = \mathrm{E}(e^{uV_i}|T_i = t) = e^{|t|}(1 + e^{-|t|}) \sum_{k=1}^{+\infty} (-1)^{k+1} k^2 \frac{\exp\left(-|t|\sqrt{k^2 - 2u}\right)}{\sqrt{k^2 - 2u}},$$

$$\mathrm{E}(V_i|T_i = t) = (1 + e^{-|t|}) \left(|t| + (1 + e^{|t|}) \log(1 + e^{-|t|})\right).$$

Thus we set

- $\alpha = \frac{1}{2}\left(1 + \frac{\pi^2 + t^2}{\mathrm{E}(V_i|T_i = t)}\right)$ to match expected values between target and proposal.
- $v^\star = 1.9834$.

The case of a Gaussian prior (I)

A special, although important, case is when the prior is $\beta \sim N(\xi, \Omega)$. In this case

- $W_i = 1, \quad i = 1, 2, \ldots, p$.
- if one set V to some specific value, say V^*, then the posterior reduces to a SUN density.
- Our strategy is to adopt a importance sampling approach in order to completely avoid the use of an MCMC scheme.
- The importance density is based on an Inverse Gamma $(\nu/2, \nu/2)$ mixture of SUN distributions that we call Unified Skew-t (SUT).
- The SUT density is obtained by fixing $V_i = V_i^*$ $(i = 1, \ldots, n)$ at a specific value and translated by an amount ξ_i^*.[2]

[2] For some values of the parameters, it belongs to the SUT family as densities as defined in Jamalizadeh and Balakrishnan (2010).

The case of a Gaussian prior (II)

The performance of the method heavily depends on an accurate choice of (ν, V^*, ξ^*).

We set ν to a large value and ξ^* to match the modes of the posterior density and the importance density.

For V^*, we have tested two methods:

1. The simplest: Take $V^* = E(V) = \pi^2/3$. [3]
2. More efficient: V^* is *optimally* chosen by *taking* $V^* = \mathrm{E}(V|\beta = \beta_{MaP})$ that is easy to compute using the close formula for $\mathrm{E}(V_i|T_i)$.

[3] It is the σ^2 value that minimises the Kullback-Leibler (KL) divergence of a centered Gaussian from a standard logistic density.

The case of a Gaussian prior (III)

Proposition 3

In an LSBR model, if the posterior distribution is a $\mathrm{pSUN}_{p,n}(Q_V, I_n, A, b, \mathbb{I}_{\{w \geq 1_p\}}, \xi, \Omega)$ *and a* $\mathrm{SUT}_{p,n}(\nu_n, I_n, \mathrm{diag}^{-\frac{1}{2}}(V_{fix})A, \mathrm{diag}^{-\frac{1}{2}}(V_{fix})b, \xi^\star, \Omega)$ *is used as the importance density, then the importance weights are bounded above for every choice of* $\xi^\star$, $\nu_n < +\infty$, V_{fix}, *and for every sample size* n.

This is sufficient to prove that the importance sampling estimator for computing the normalising constant of the posterior has finite variance.

Prior Distributions for Logit Model

We have adopted a pSUN prior with weakly informative values of the hyperparameters in the spirit of Gelman et al. (2008).

We set $m = 0, \xi = 0_p$ and a diagonal matrix for Ω. We consider three different priors: Gaussian, Laplace with independent components (*Laplacit*), and a Dirichlet-Laplace prior.

The diagonal components of Ω for the Gaussian prior are obtained by multiplying the ones of Durante (2019) by $\pi^2/3$, i.e. the mean of a logistic Kolmogorov distribution.

In the other cases, we have set the scaling parameters in order to have roughly the same prior credible interval at level 95% of the Gaussian case, thus

$$\begin{aligned} \text{Gaussian:} \quad & \omega_{1,1} = \cdots = \omega_{pp} = 52.6379\,; \\ \text{Laplacit:} \quad & \omega_{1,1} = \cdots = \omega_{pp} = 22.5314\,; \\ \text{Dirichlet-Laplace:} \quad & \omega_{1,1} = \cdots = \omega_{pp} = 9.0984\,. \end{aligned}$$

Polya-Gamma vs pSUN with Small Sample Size (I)

- We evaluate the performance of Polya-Gamma (PG) Gibbs sampler, Ultimate Polya-Gamma (UPG) Gibbs sampler (Zens et al., 2023), pSUN Gibbs sampler, and pSUN importance sampling.
- We set the sample size $n = 50$, the number of parameters $p = 500$, the number of draws from the posterior $N = 10^4$ ($N = 10^5$ for the case of Dirichlet-Laplace prior); we then iterate the entire simulation $G = 2400$ times.

Polya-Gamma vs pSUN with Small Sample Size (II)

We keep fixed the true parameter vector $\beta^\dagger$ at the following values:

$$\begin{aligned}
\beta_1^\dagger &= \Lambda^{-1}(0.50) = 0\,, & \beta_2^\dagger, \ldots, \beta_{26}^\dagger &= \Lambda^{-1}(0.05) = -2.9444\,,\\
\beta_{27}^\dagger, \ldots, \beta_{76}^\dagger &= \Lambda^{-1}(0.10) = -2.1972, & \beta_{77}^\dagger, \ldots, \beta_{126}^\dagger &= \Lambda^{-1}(0.20) = -1.3863,\\
\beta_{127}^\dagger, \ldots, \beta_{176}^\dagger &= \Lambda^{-1}(0.30) = -0.8473, & \beta_{177}^\dagger, \ldots, \beta_{226}^\dagger &= \Lambda^{-1}(0.40) = -0.4055,\\
\beta_{227}^\dagger, \ldots, \beta_{276}^\dagger &= \Lambda^{-1}(0.60) = 0.4055, & \beta_{277}^\dagger, \ldots, \beta_{326}^\dagger &= \Lambda^{-1}(0.70) = 0.8473,\\
\beta_{327}^\dagger, \ldots, \beta_{376}^\dagger &= \Lambda^{-1}(0.80) = 1.3863, & \beta_{377}^\dagger, \ldots, \beta_{426}^\dagger &= \Lambda^{-1}(0.90) = 2.1972,\\
\beta_{427}^\dagger, \ldots, \beta_{451}^\dagger &= \Lambda^{-1}(0.95) = 2.9444, & \beta_{452}^\dagger, \ldots, \beta_{500}^\dagger &= \Lambda^{-1}(0.50) = 0,
\end{aligned}$$

where $\Lambda(\cdot)$ is the CDF of the standard logistic distribution.

Polya-Gamma vs pSUN with Small Sample Size (III)

We run the following simulation scheme

set $X_{i,1} = 1$, $i = 1, 2, \ldots, n$

for $g = 1, 2, \ldots, G$

- sample $X_{ij} \overset{\text{iid}}{\sim} N(0, 1)$, $i = 1, 2, \ldots, n$, $j = 2, 3, \ldots, p$
- center and scale each column of X, except the first, in order to have a standard deviation equal to 0.5
- sample $Y_i \overset{\text{ind}}{\sim} Be\Big(\Lambda(X_i' \beta^{\dagger})\Big)$, $i = 1, 2, \ldots, n$
- draw N values from the posterior distribution of β

$\Longrightarrow$ compute $\Big(\widehat{\mathrm{E}}(\beta|Y) - \beta^{\dagger}\Big)$, ESS, ACF (only for MCMC) and get the computational time

Polya-Gamma vs pSUN with Small Sample Size: ESS per second

Polya-Gamma vs pSUN: Gaussian Prior. Mean of ESS per second for the Different Groups of β

Gaussian	UPG	PG	pSUN-Gibbs	pSUN-IS
Intercept	0.42	1.51	44.34	374.41
$\beta = \Lambda^{-1}(0.05)$	10.57	19.29	44.32	374.41
$\beta = \Lambda^{-1}(0.10)$	10.55	19.26	44.32	374.41
$\beta = \Lambda^{-1}(0.20)$	10.53	19.28	44.31	374.41
$\beta = \Lambda^{-1}(0.30)$	10.52	19.25	44.31	374.41
$\beta = \Lambda^{-1}(0.40)$	10.56	19.27	44.32	374.41
$\beta = \Lambda^{-1}(0.50)$	10.55	19.27	44.32	374.41
$\beta = \Lambda^{-1}(0.60)$	10.53	19.26	44.32	374.41
$\beta = \Lambda^{-1}(0.70)$	10.56	19.28	44.32	374.41
$\beta = \Lambda^{-1}(0.80)$	10.55	19.25	44.32	374.41
$\beta = \Lambda^{-1}(0.90)$	10.54	19.27	44.31	374.41
$\beta = \Lambda^{-1}(0.95)$	10.59	19.25	44.33	374.41

Polya-Gamma vs pSUN with Small Sample Size: ESS per second

Polya-Gamma vs pSUN: Mean of ESS Per Second with Laplacit (Left) and Dirichlet-Laplace (Right) Priors for the Different Groups of β

Laplacit	UPG	PG	pSUN-Gibbs
Intercept	0.36	1.30	24.80
$\beta = \Lambda^{-1}(0.05)$	7.30	13.89	26.15
$\beta = \Lambda^{-1}(0.10)$	7.34	13.97	26.35
$\beta = \Lambda^{-1}(0.20)$	7.34	14.06	26.52
$\beta = \Lambda^{-1}(0.30)$	7.38	14.09	26.58
$\beta = \Lambda^{-1}(0.40)$	7.35	14.08	26.60
$\beta = \Lambda^{-1}(0.50)$	7.39	14.10	26.64
$\beta = \Lambda^{-1}(0.60)$	7.38	14.10	26.61
$\beta = \Lambda^{-1}(0.70)$	7.39	14.09	26.58
$\beta = \Lambda^{-1}(0.80)$	7.36	14.04	26.50
$\beta = \Lambda^{-1}(0.90)$	7.34	13.97	26.36
$\beta = \Lambda^{-1}(0.95)$	7.32	13.86	26.12

Dirichlet-Laplace	UPG	PG	pSUN-Gibbs
Intercept	0.78	1.95	11.12
$\beta = \Lambda^{-1}(0.05)$	0.46	1.23	7.83
$\beta = \Lambda^{-1}(0.10)$	0.48	1.27	8.06
$\beta = \Lambda^{-1}(0.20)$	0.48	1.29	8.21
$\beta = \Lambda^{-1}(0.30)$	0.49	1.31	8.28
$\beta = \Lambda^{-1}(0.40)$	0.49	1.31	8.32
$\beta = \Lambda^{-1}(0.50)$	0.49	1.32	8.33
$\beta = \Lambda^{-1}(0.60)$	0.49	1.32	8.34
$\beta = \Lambda^{-1}(0.70)$	0.49	1.31	8.31
$\beta = \Lambda^{-1}(0.80)$	0.49	1.30	8.24
$\beta = \Lambda^{-1}(0.90)$	0.47	1.27	8.05
$\beta = \Lambda^{-1}(0.95)$	0.46	1.24	7.87

Polya-Gamma vs pSUN with Small Sample Size: MSE

However, the Dirichlet-Laplace prior typically shows a better MSE

Polya-Gamma vs pSUN: MSE for the Different Groups of β and Priors with pSUN approaches

	pSUN-Gibbs			pSUN-IS
	Gaussian	**Laplacit**	**Dirichlet-Laplace**	**Gaussian**
Intercept	13.38	8.34	0.41	13.21
$\beta = \Lambda^{-1}(0.05)$	7.43	7.40	8.82	7.42
$\beta = \Lambda^{-1}(0.10)$	5.47	5.24	5.25	5.45
$\beta = \Lambda^{-1}(0.20)$	4.05	3.65	2.69	4.02
$\beta = \Lambda^{-1}(0.30)$	3.43	2.99	1.55	3.40
$\beta = \Lambda^{-1}(0.40)$	3.15	2.71	1.07	3.12
$\beta = \Lambda^{-1}(0.50)$	3.08	2.59	0.92	3.05
$\beta = \Lambda^{-1}(0.60)$	3.12	2.70	1.12	3.09
$\beta = \Lambda^{-1}(0.70)$	3.47	3.00	1.65	3.43
$\beta = \Lambda^{-1}(0.80)$	4.03	3.66	2.66	4.00
$\beta = \Lambda^{-1}(0.90)$	5.48	5.25	5.24	5.46
$\beta = \Lambda^{-1}(0.95)$	7.44	7.34	8.52	7.43

Cancer SAGE Example (I)

(Skip this)

Discussed in Durante (2019) a $p > n$ case: $n = 74$ observations with 516 covariates.

Of interest: to quantify the effects of gene expressions on the probability of a cancerous tissue and to predict the status of new tissues as a function of the gene expression.

Cancer SAGE Example (II)

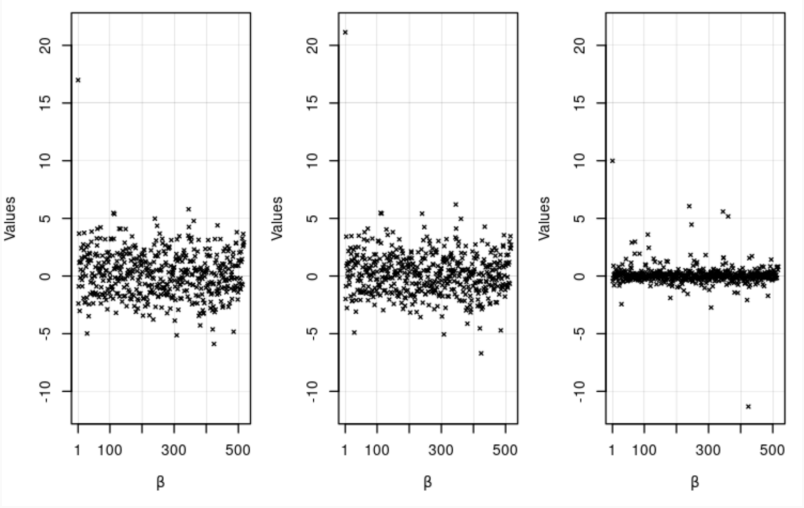
Regarding the effective sample size per second, the results essentially confirm the findings of the simulation study.

Cancer SAGE Example: Summaries of Effective Sample Size Per Second for the Different Priors of β and Algorithms

Algorithm	Prior	Min	1st Qu.	Median	Mean	3rd Qu.	Max
UPG	Gaussian	0.21	1.75	2.55	3.13	3.90	16.55
PG	Gaussian	1.12	12.46	14.40	14.93	16.83	30.12
pSUN-Gibbs	Gaussian	29.67	30.91	30.91	30.90	30.91	33.26
pSUN-IS	Gaussian	209.81	209.81	209.81	209.81	209.81	209.81
UPG	Laplacit	0.08	1.35	1.87	2.26	2.76	9.96
PG	Laplacit	0.43	8.21	9.62	9.92	11.42	20.40
pSUN-Gibbs	Laplacit	9.36	18.83	20.83	20.05	22.04	23.28
UPG	Dirichlet-Laplace	0.03	0.19	0.38	0.47	0.69	2.13
PG	Dirichlet-Laplace	0.06	0.46	0.99	1.24	1.86	5.57
pSUN-Gibbs	Dirichlet-Laplace	0.08	0.67	1.13	1.21	1.71	2.70

Cancer SAGE Example (III)

Posterior means for different priors: Gaussian (Left), Laplace (Center), and Dirichlet-Laplace (Right)



Lung Cancer Example (I)

Discussed in Hong and Yang (1991), $n = 32$ observations with $p = 55$ covariates.

Of interest: variable selection and link comparison (logit vs. probit).

Lung Cancer Example (II)

We computed – both in the logit and the probit cases – the median probability model, that is, the subset of covariates such that their posterior inclusion probability is larger than 0.5. (Barbieri and Berger, 2004)

The final models are

- logit model with 38 variables
- probit model with 29 variables
- 18 covariates are common to both median models

An interesting by-product of our approach is that one can also make a model comparison between logit and probit links.

Variable Selection

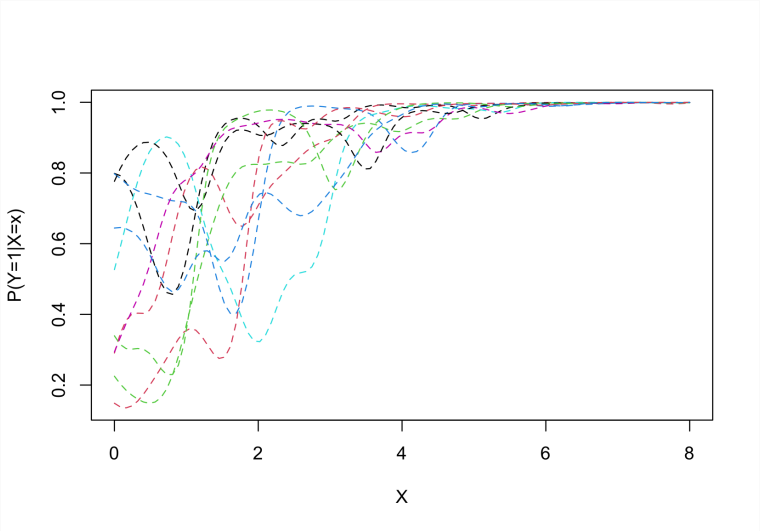
We assumed a uniform prior over the set of models and compatible Gaussian priors within each model, both in the logit and in the probit cases.

We use a Gibbs variable selection algorithm to obtain a posterior sample from the space of all possible subsets of covariates.

The Bayes factor of the median logit model vs. the median probit one is

$$\mathrm{BF}_{L,P} = \frac{3.72 \times 10^{-8}}{3.27 \times 10^{-9}} = 11.38.$$

PART II: Nonparametrics



A Nonparametric Approach

We have only considered the standard linear calibration function $\eta(x) = x'\beta$. In some cases, in the presence of many observations and/or many covariates, this assumption might be too restrictive.

What happens if we assume a nonparametric calibration function? The new model is then a Nonparametric Symmetric Binary Regression (NSBR) model.

However, we cannot use the pSUN theory in a nonparametric framework, because of its lack of closeness w.r.t. to conditioning, marginalisation and affine transformations.

The qSUN Family of Distributions

A random vector η has a *Quasi* SUN (qSUN) distribution, denoted

$$\eta \sim \mathrm{qSUN}_{d,m}(Q_V, \Theta, \Psi, A, b, \xi, \Omega)$$

if the following stochastic representation holds

$$\begin{aligned}
\eta &= \xi + \mathrm{diag}^{\frac{1}{2}}(\Omega) Z \,|\, T + S \leq AZ + b\,, \\
T|V &\sim N_m(0, \Theta_V)\,, \\
S &\sim N_m(0, \Psi)\,, \\
Z &\sim N_d(0, \bar{\Omega})\,, \\
V &\sim Q_V(\cdot)\,,
\end{aligned}$$

where $\Theta_V = \mathrm{diag}^{1/2}(V)\,\Theta\,\mathrm{diag}^{1/2}(V)$, Θ is a correlation matrix, Ψ is a generic positive semi-definite matrix, and all variables are mutually independent, excluding, of course, V and T.

Remark: W_i's all equal to 1.

qSUN PDF and MGF

Set $\Upsilon_{Q_V,\Theta,\Psi}(b) = \mathrm{P}(T + S \leq b)$, $T \sim \Phi_{Q_V,\Theta} \perp\!\!\!\perp S \sim N_m(0_m, \Psi)$.

If $\eta \sim \mathrm{qSUN}_{d,m}(Q_V, \Theta, \Psi, A, b, \xi, \Omega)$ then

$$p_\eta(\eta) = \varphi_\Omega(\eta - \xi) \frac{\Upsilon_{Q_V,\Theta,\Psi}\left(A\mathrm{diag}^{-\frac{1}{2}}(\Omega)(\eta - \xi) + b\right)}{\Upsilon_{Q_V,\Theta,\Psi+A\bar{\Omega}A'}(b)},$$

$$M_\eta(u) = \exp\left(u'\xi + \frac{1}{2}u'\Omega u\right) \frac{\Upsilon_{Q_V,\Theta,\Psi+A\bar{\Omega}A'}\left(A\bar{\Omega}\mathrm{diag}^{\frac{1}{2}}(\Omega)u + b\right)}{\Upsilon_{Q_V,\Theta,\Psi+A\bar{\Omega}A'}(b)}.$$

Comparisons

1. SUN:

$$\beta = \xi + \text{diag}^{\frac{1}{2}}(\Omega) Z \mid T \leq AZ + b, \text{ with } A \in \mathbb{R}^{m \times p}, b \in \mathbb{R}^m.$$

2. pSUN: Assume that $Z|W \sim N_p(0_p, \bar{\Omega}_W)$, $T|V \sim N_m(0_m, \Theta_V)$ with

$$V \sim Q_V(\cdot) \quad \perp\!\!\!\perp \quad W \sim Q_W(\cdot)$$

The pSUN class is defined as the expression above with the new assumptions on Z and T. Then,

$$\text{pSUN}_{p,m}(Q_V, \Theta, A, b, Q_W, \xi, \Omega)$$

3. qSUN:

$$\eta = \xi + \text{diag}^{\frac{1}{2}}(\Omega) Z | T + S \leq AZ + b,$$
$$T|V \sim N_m(0, \Theta_V),\ S \sim N_m(0, \Psi),\ Z \sim N_d(0, \bar{\Omega}),\ V \sim Q_V(\cdot),$$

and all variables are mutually independent, except for V and T. W_i's are all equal to 1.

Closure Properties

Using the moment generating function of a qSUN distribution, it is possible to show that if

$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \sim \mathrm{qSUN}_{d_1+d_2,m}\left(Q_V, \Theta, \Psi, \begin{bmatrix} A_1 & A_2 \end{bmatrix}, b, \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}, \begin{bmatrix} \Omega_{1,1} & \Omega_{1,2} \\ \Omega'_{1,2} & \Omega_{2,2} \end{bmatrix}\right)$$

then

- $\eta_1 \sim \mathrm{qSUN}_{d_1,m}(Q_V, \Theta, \Theta_2^*, A^*, b, \xi_1, \Omega_{1,1})$,
- $\eta_1 | \eta_2 = x \sim \mathrm{qSUN}_{d_1,m}(Q_V, \Theta, \Psi, \widetilde{A}, \widetilde{b}, \widetilde{\xi}, \widetilde{\Omega})$,
- let $p \in \mathbb{R}^s$ and $P \in \mathbb{R}^{s \times (d_1+d_2)}$, thus $p + P\eta \sim \mathrm{qSUN}_{s,m}(Q_V, \Theta, \Psi_P, A_P, b, \xi_{P,p}, \Omega_P)$.

Remark: Taking only the W_i's all equal to 1 would be enough for conditioning but not for marginalisation.

qSUN Process

Recall:

$$\eta = \xi + \mathrm{diag}^{\frac{1}{2}}(\Omega) Z | T + S \leq AZ + b\,,$$

Set $R = AZ - S$, then

$$R | T + S \leq AZ + b \sim \mathrm{qSUN}_{m,m}(Q_V, \Theta, 0_{m\times m}, \mathrm{diag}^{\frac{1}{2}}(C), b, 0_m, C) \text{ with } C = A\bar{\Omega}A' + \Psi.$$

We construct the qSUN process assuming that this variable R is common to all the variables in the process.

The qSUN distribution is not identifiable without a restriction on Q_V.

The qSUN Process

We say $\eta(x)$ for $x \in \mathbb{R}^p$ is a qSUN process if

$$\begin{aligned}
\eta(x) &= m(x) + \sqrt{k(x,x)}\,\zeta(x) \mid T \leq R + b \\
T \mid V &\sim \; N_m(0, \Theta_V)\,,\; V \sim Q_V(\cdot)\,, \\
R &\sim N_m(0_m, C)\,, \\
\zeta(x) &\sim \mathrm{GP}(0, \bar{k}(x,x'))\,, \\
\nu(x) &= \mathrm{Cov}\big(R, \zeta(x)\big) \\
\bar{k}(x,x') &= k(x,x')/\sqrt{k(x,x)k(x',x')}\,.
\end{aligned}$$

We denote $\eta(x) \sim \mathrm{qSUNpro}_m(Q_V, \Theta, C, b, \nu(x), m(x), k(x,x'))$.

The functions $\nu(\cdot)$, $\bar{k}(\cdot,\cdot)$, and the matrix C must satisfy that

$$\begin{bmatrix} C & \nu(X) \\ \nu'(X) & \bar{k}(X,X) \end{bmatrix} \text{ is positive semi-definite } \forall X \in \mathbb{R}^{d \times p}$$

This is an extension of the $\mathrm{Skew\text{-}GP}$ process of Benavoli et al. (2020) to include the logit case.

Distribution of Every Finite Collection of Points

The following Proposition shows that every finite collection of points from a qSUN process follows a qSUN distribution.

Proposition 4
If $\eta(x) \sim \mathrm{qSUNpro}_m(Q_V, \Theta, C, b, \nu(x), m(x), k(x, x'))$, *then* $\forall X \in \mathbb{R}^{d \times p}$ *the following holds*

$$\eta(X) \sim \mathrm{qSUN}_{d,m}(Q_V, \Theta, \Psi(X), A(X), b, m(X), k(X, X))\,,$$

where

$$\Psi(X) = C - \nu(X)\bar{k}^{-1}(X, X)\nu'(X)\,,\ A(X) = \nu(X)\bar{k}^{-1}(X, X)$$

Conditional Process

The process is also closed under conditioning.

Proposition 5

If $\eta(x) \sim \mathrm{qSUNpro}_m(Q_V, \Theta, C, b, \nu(x), m(x), k(x,x'))$, *then*

$$\eta(x)|\eta(X_H) \sim \mathrm{qSUNpro}_m(Q_V, \Theta, C^\star, b^\star, \nu^\star(x), m^\star(x), k^\star(x,x'))\,, \forall X_H \in \mathbb{R}^{h\times p}\,,$$

where

$$\begin{aligned}
C^\star &= C - \nu(X_H)\bar{k}^{-1}(X_H, X_H)\nu'(X_H)\,,\\
b^\star &= b + \nu(X_H)\bar{k}^{-1}(X_H, X_H)\mathrm{diag}^{-\frac{1}{2}}\big(k(X_H, X_H)\big)\big(\eta(X_H) - m(X_H)\big)\,,\\
\nu^\star(x) &= \sqrt{\frac{k(x,x)}{k^\star(x,x)}}\Big(\nu(x) - \nu(X_H)\bar{k}^{-1}(X_H, X_H)\bar{k}(X_H, x)\Big)\,,\\
m^\star(x) &= m(x) + k(x, X_H)k^{-1}(X_H, X_H)\big(\eta(X_H) - m(X_H)\big)\,,\\
k^\star(x,x') &= k(x,x') - k(x, X_H)k^{-1}(X_H, X_H)k(X_H, x)\,.
\end{aligned}$$

Sum of q i.i.d. qSUN Processes

The sum of q i.i.d. qSUN processes is still a qSUN process.

Proposition 6

If $\eta(x) = \sum_{i=1}^{q} \eta_i(x)$ *and* $\eta_i(x) \overset{i.i.d.}{\sim} \mathrm{qSUNpro}_{m_i}(Q_{V_i}, \Theta_i, C_i, b_i, \nu_i(x), m_i(x), k_i(x,x'))$*, then*

$$\eta(x) \sim \mathrm{qSUNpro}_{m}(Q_V, \Theta, C, b, \nu(x), m(x), k(x,x'))\,,$$

where

$$m = \sum_{i=1}^{q} m_i\,,\ m(x) = \sum_{i=1}^{q} m_i(x)\,,\ k(x,x') = \sum_{i=1}^{q} k_i(x,x')\,,\ Q_V(v) = \prod_{i=1}^{q} Q_{V_i}(v_i)\,,$$

$$\Theta = \begin{bmatrix} \Theta_1 & & \\ & \ddots & \\ & & \Theta_q \end{bmatrix},\ C = \begin{bmatrix} C_1 & & \\ & \ddots & \\ & & C_q \end{bmatrix},\ \nu(x) = \frac{1}{\sqrt{k(x,x)}} \begin{bmatrix} \sqrt{k_1(x,x)}\nu_1(x) \\ \cdots \\ \sqrt{k_q(x,x)}\nu_q(x) \end{bmatrix}.$$

Admissible Forms for $C, \nu(x)$ depending on $k(x, x')$ (I)

Case 1: Take $R = A\zeta(X_H) - S$, $S \sim N_m(0_m, \Psi) \perp\!\!\!\perp \zeta(x)$ with $X_H \in \mathbb{R}^{h \times p}$.

In this case, it is easy to show that:

$$C = \Psi + A\bar{k}(X_H, X_H)A', \; \nu(x) = A\bar{k}(X_H, x),$$
$$\eta(X_H) \sim \mathrm{qSUN}_{h,m}(Q_V, \Theta, \Psi, A, b, m(X_H), k(X_H, X_H)),$$

but using Proposition 5 is simple to see that

$$\eta(x)|\eta(X_H) \sim \mathrm{GP}(m^\star(x), k^\star(x, x')).$$

To avoid a conditional Gaussian process, we combine Case 1 and Proposition 6 in Case 2.

Admissible Forms for $C, \nu(x)$ depending on $k(x, x')$ (II)

Case 2: Take $\eta(x) = \sum_{i=1}^{q} \eta_i(x)$, $\eta_i(x) \overset{i.i.d.}{\sim} \mathrm{qSUNpro}_{m_i}(Q_{V_i}, \Theta_i, C_i, \nu_i(x), m_i(x), k_i(x, x'))$, and $C_i = \Psi_i + A_i \bar{k}_i(X_{H_i}, X_{H_i}) A_i'$, $\nu_i(x) = A_i \bar{k}(X_{H_i}, x)$.

Using Proposition 6 is easy to show that

$$\eta(x) \sim \mathrm{qSUNpro}_m(Q_V, \Theta, C, b, \nu(x), m(x), k(x, x'))\,,$$

where

$$m = \sum_{i=1}^{q} m_i\,,\ m(x) = \sum_{i=1}^{q} m_i(x)\,,\ k(x, x') = \sum_{i=1}^{q} k_i(x, x')\,,\ Q_V(v) = \prod_{i=1}^{q} Q_{V_i}(v_i)\,,$$

$$\Theta = \begin{bmatrix} \Theta_1 & & \\ & \ddots & \\ & & \Theta_q \end{bmatrix},\ C = \begin{bmatrix} \Psi_1 + A_i \bar{k}_1(X_{H_1}, X_{H_1}) A_1' & & \\ & \ddots & \\ & & \Psi_q + A_q \bar{k}_1(X_{H_q}, X_{H_q}) A_q' \end{bmatrix},$$

$$\nu(x) = \frac{1}{\sqrt{k(x, x)}} \begin{bmatrix} \sqrt{k_1(x, x)} A_1 \bar{k}(X_{H_1}, x) \\ \cdots \\ \sqrt{k_q(x, x)} A_q \bar{k}(X_{H_q}, x) \end{bmatrix}.$$

In words

The new process $\eta(x)$ is essentially a Gaussian process perturbed in h points, say X_H, through m Gaussian latent variables.

Due to the perturbation, the qSUN process is not stationary even using a stationary covariance function $k(x, x')$.

Conjugacy for Nonparametric Symmetric Binary Regression (NSBR)

Proposition 7

In a Bayesian NSBR model, if $\eta(x) \sim \mathrm{qSUNpro}_m(Q_V, \Theta, \Psi, C, b, \nu(x), m(x), k(x,x'))$ *and the link function is of the form* $\Lambda(x) = \int_0^\infty \Phi\left(\frac{x}{\sqrt{v}}\right) dQ_{V_0}(v)$*, then the following holds:*

$$\eta(x)|Y = y \sim \mathrm{qSUNpro}_{m+n}\left(Q_V^\star, \Theta^\star, C^\star, b^\star, \nu^\star(x), m(x), k(x,x')\right),$$

where

$$Q^\star([v_1 v_2]') = Q_V(v_1) \prod_{i=1}^{n} Q_{V_0}(v_{2,i}),$$

$$\Theta^\star = \begin{bmatrix} \Theta & 0_{m\times n} \\ 0_{n\times m} & I_n \end{bmatrix}, \; C^\star = \begin{bmatrix} C & \nu(X)\mathrm{diag}^{-1/2}\left(k(X,X)\right)B_y \\ B_y\mathrm{diag}^{\frac{1}{2}}\left(k(X,X)\right)\nu'(X) & B_y, k(X,X)B_y \end{bmatrix},$$

$$b^\star = \begin{bmatrix} b \\ B_y m(X) \end{bmatrix}, \; \nu^\star(x) = \begin{bmatrix} \nu(x) \\ B_y\mathrm{diag}^{\frac{1}{2}}\left(k(X,X)\right)\bar{k}(X,x) \end{bmatrix}, \; B_y = \begin{bmatrix} 2y_1 - 1 & & \\ & \ddots & \\ & & 2y_n - 1 \end{bmatrix}$$

Posterior Computation

The stochastic representation of a qSUN process allows the use of popular algorithms:

- MCMC: generalised elliptical slice sampling to improve the mixing of the popular elliptical slice sampling (Murray et al., 2009).
- EP: quasi-Monte Carlo algorithm for performing the Expectation-Propagation method.
- VB: with a further constraint in the mean field[4], the update in the CAVI algorithm is available in closed form.

[4] independence among the R_i's

Conclusions

Binary regression is one of the most popular tools in applied statistics. The availability of a posterior sampler is important for producing accurate and ready-to-use summaries.

1. We have proposed an algorithm which – in the special logit regression case with a Gaussian prior – produces a weighted posterior sample of independent draws.
   This allows easy computation of posterior summaries, including the normalizing constant, crucial to perform model selection.
2. When both the link function and the prior are scale mixtures of Gaussian densities, the method uses a Gibbs approach with a uniformly better mixing than competitive algorithms, with a dramatic difference in the $p > n$ case
3. Our approach can be extended to cover the case of a nonlinear predictor, through the definition of a *quasi*-SUN process (work in progress, almost finished).

References

D. F. Andrews and C. L. Mallows. Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102, 1974.

R. B. Arellano-Valle and A. Azzalini. On the unification of families of skew-normal distributions. *Scand. J. Statist.*, 33(3): 561–574, 2006.

M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32:870–897, 2004.

O. Barndorff-Nielsen. Exponentially decreasing distributions for the logarithm of particle size. In *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, , pages 401–409. The Royal Society, London, 1977.

A. Benavoli, D. Azzimonti, and D. Piga. Skew gaussian processes for classification. *Machine Learning*, 109:1877–1902, 2020.

Z. I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(1):125–148, 2017.

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

A. Gelman, A. Jakulin, M. G. Pittau, and Y. Su. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.*, 2(4), 2008.

References ii

C. C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.*, 1(1): 145–168, 2006.

Z.Q. Hong and J.Y. Yang. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. *Pattern Recognition*, 24(4):317–324, 1991.

A. Jamalizadeh and N. Balakrishnan. Distributions of order statistics and linear combinations of order statistics from an elliptical distribution as mixtures of unified skew-elliptical distributions. *Journal of Multivariate Analysis*, 101(6):1412–1427, 2010.

I. Murray, R. P. Adams, and D. J. C. MacKay. Elliptical slice sampling. In *International Conference on Artificial Intelligence and Statistics*, 2009.

P. Onorati and B. Liseo. Random number generator for the Kolmogorov distribution. *arXiv2208.13598*, 2022.

P. Onorati and B. Liseo. An extension of the unified skew-normal family of distributions and its application to Bayesian binary regression. *Journal of Computational and Graphical Statistics*, 34(4):1291–1304, 2025.

D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.*, 4(4):673–684, 1976.

L. A. Stefanski. A normal scale mixture representation of the logistic distribution. *Statistics & Probability Letters*, 11(1):69–70, 1991.

G. Zens, S. Frühwirth-Schnatter, and H. Wagner. Ultimate Pólya Gamma Samplers–Efficient MCMC for Possibly Imbalanced Binary and Categorical Data. *Journal of the American Statistical Association*, 0(0):1–12, 2023.

The Density Function and Moment Generating Function of SUN

The density function of a SUN vector includes the computation of two CDFs of a multivariate Gaussian density

$$f_\beta(\beta) = \varphi_\Omega(\beta - \xi)\frac{\Phi_{\Gamma - \Delta'\bar{\Omega}^{-1}\Delta}(\tau + \Delta'\bar{\Omega}^{-1}\mathrm{diag}^{-\frac{1}{2}}(\Omega)(\beta - \xi))}{\Phi_\Gamma(\tau)}.$$

It is also available the moment generating function

$$M_\beta(u) = \mathrm{E}(e^{u'\beta}) = e^{u' + u'\Omega u/2}\frac{\Phi_\Gamma(\tau + \Delta'\mathrm{diag}^{\frac{1}{2}}(\Omega)u)}{\Phi_\Gamma(\tau)}.$$

The MGF of a pSUN

Assume $M_Z(u)$ (MGF of Z) exists. Then, the MGF of β is

$$M_Y(u) = e^{u'\xi} M_Z\left(\text{diag}^{\frac{1}{2}}(\Omega)u\right) \frac{\widetilde{\Psi}_{Q_V,\Theta,A,Q_W,\bar{\Omega}}\left(b, \text{diag}^{\frac{1}{2}}(\Omega)u\right)}{\Psi_{Q_V,\Theta,A,Q_W,\bar{\Omega}}(b)},$$

with

$$\widetilde{\Psi}_{Q_V,\Theta,A,b,Q_W,\bar{\Omega}}(b,k) = \mathrm{P}(T - A\widetilde{Z}_k \leq b)$$

$T \sim \Phi_{\Theta,Q_V}(\cdot) \perp\!\!\!\perp \widetilde{Z}_k$, and $\widetilde{Z}_k$ is the k-tilted distribution Siegmund (1976) of $Z \sim \Phi_{\bar{\Omega},Q_W}(\cdot)$

pSUN: Conditional and Marginal (I)

Consider the following partition:

$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \in \mathbb{R}^{d_1+d_2},$$

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \operatorname{diag}^{\frac{1}{2}} \left(\begin{bmatrix} \Omega_{1,1} & \Omega_{1,2} \\ \Omega_{1,2}' & \Omega_{2,2} \end{bmatrix} \right) \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \Bigg| \left(T \leq \begin{bmatrix} A_1 & A_2 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + b \right),$$

$$Z_1, Z_2 | W_1, W_2 \sim N_{d_1+d_2} \left(0_{d_1+d_2}, \begin{bmatrix} \operatorname{diag}^{\frac{1}{2}}(W_1) & 0_{d_1 \times d_2} \\ 0_{d_2 \times d_1} & \operatorname{diag}^{\frac{1}{2}}(W_2) \end{bmatrix} \bar{\Omega} \begin{bmatrix} \operatorname{diag}^{\frac{1}{2}}(W_1) & 0_{d_1 \times d_2} \\ 0_{d_2 \times d_1} & \operatorname{diag}^{\frac{1}{2}}(W_2) \end{bmatrix} \right),$$

$$\begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \sim Q_{W_1, W_2}(\cdot).$$

pSUN: Conditional and Marginal (II)

$\eta_1|\eta_2 = x$ and η_1 are not pSUN random variables.

The stochastic representations of these involve random matrices depending on (W_1, W_2).

The idea is to set (W_1, W_2) equal to a vector of ones.

In this way we obtain a random variable that is closed under conditioning. To obtain closeness w.r.t. marginalisation, one has to split T into two components.

Pulsar Stars Example (I)

Pulsar Stars dataset: $n = 17\,898$ observations with $p = 8$ covariates.

Of interest: to distinguish whether an electromagnetic radiation comes from a pulsar star or from another source.

Pulsar Stars Example (II)

We adopt a 2-fold cross-validation strategy.

Performance Measurement for the Pulsar Star Dataset

	Accuracy	Sensitivity	Specificity	Precision	F1 Score
First Group	0.9803	0.8688	0.9916	0.9132	0.8904
Second Group	0.9779	0.8284	0.9929	0.9210	0.8723
Mean	0.9791	0.8486	0.9923	0.9171	0.8813

Approximate Bayes (VB 1)

- Let $\pi(\theta \mid X)$ be the intractable posterior distribution and let $q(\theta)$ be a density belonging to $\mathcal{Q}$, a general class of tractable densities. An optimal approximation $\hat{q}(\theta) \in \mathcal{Q}$ of the posterior distribution is defined as

$$\hat{q}(\theta) = \arg\min_{q \in \mathcal{Q}} \mathcal{D}\big(q(\theta), \pi(\theta \mid X)\big)$$

- where $\mathcal{D}(\cdot, \cdot)$ is some divergence or metric over the space of probability distributions.
- An example is the Kullback–Leibler divergence of g from f: $KL(f \| g) = \int_\Theta f(\theta) \log \frac{f(\theta)}{g(\theta)} \, d\theta$
- Depending on the choice of $\mathcal{D}$ and the class $\mathcal{Q}$, the problem can be computationally feasible or not.
- Of course, the actual posterior $\pi(\theta \mid X)$ should not be included in $\mathcal{Q}$, otherwise

$$\hat{q}(\theta) = \pi(\theta \mid X),$$

for almost all reasonable divergence measures $\mathcal{D}$.

Approximate Bayes (VB 2)

- As for the choice of $\mathcal{D}(\cdot, \cdot)$, it would be theoretically appealing to consider metrics such as the Hellinger distance, the total variation distance, or the Wasserstein distance.
- Unfortunately, even when we assume $\mathcal{Q}$ be the space of multivariate Gaussians, finding the optimal density $\hat{q}(\theta)$ could be problematic.
- A basic requirement is that the optimisation procedure should not depend on the intractable normalising constant of the posterior.

As a consequence, we only consider two divergences.

- $KL\big[q(\theta) \,\|\, \pi(\theta \mid X)\big]$ leading to Variational Bayes methods
- Expectation Propagation methods

The evidence lower bound (ELBO) [1]

Define the ELBO of a density $q(\theta)$ as

$$\text{ELBO}(q(\theta)) = \int_{\Theta} q(\theta) \log \frac{\pi(\theta, X)}{q(\theta)} d\theta = -KL(q(\theta)||\pi(\theta, X))$$

It can be shown that

$$\log(\pi(X)) = KL(q(\theta)||\pi(\theta|X)) + \text{ELBO}(q(\theta))$$

Since $\log \pi(X)$ does not depend on θ, one obtains that

$$\hat{q}(\theta) = \arg\min_{q \in \mathcal{Q}} KL(q(\theta)||\pi(\theta|X)) = \arg\max_{q \in \mathcal{Q}} \text{ELBO}(q(\theta)),$$

so the optimisation will not depend on the normalising constant!

The evidence lower bound (ELBO) [2]

- ELBO is a lower bound of the value of $\log \pi(X)$ because $KL \geq 0$, which implies $\text{ELBO}(q(\theta)) \leq \log \pi(X)$
- This property has suggested using the variational bound as a model selection criterion, assuming that the ELBO will be a good approximation of $\log \pi(X)$.
- Even when an optimal distribution $\hat{q}(\theta)$ can be found, there is no guarantee that the minimised KL will be small in absolute terms.
- Essentially, it is still hard to assess the quality of the obtained approximation without comparing it with some gold standard, such as MCMC.

Mean-Field Approximation

The VB optimisation problem is ill-posed if we do not specify a tractable class $\mathcal{Q}$.

- A convenient assumption is to restrict the focus on a class $\mathcal{Q}$ of mean-field approximations, in which we assume

$$q(\theta) = \prod_{k=1}^{K} q(\theta_k)$$

  implying that we are forcing independence among the K groups of parameters.
- Dependence is preserved within each block of parameters.
- However, $q(\theta)$ is not forced to belong to any specific parametric family of distributions. The only assumption we are making is independence among the K groups

Mean Field limitations

- The mean-field family can capture any marginal density. However, it cannot capture the correlation between them.
- While the variational approximation has the same mean as the original density, its covariance structure is, by construction, decoupled.
- The marginal variances of the approximation under-represent those of the target density. (plug-in solution)
- The KL divergence penalizes placing mass in $q(\cdot)$ on areas where $\pi(\cdot)$ has little mass, but penalizes less the reverse.

Derivation of the CAVI algorithm [1]

Coordinate Ascent Variational Inference

Under the mean-field assumption, the optimisation of the ELBO can be written as

$$\text{ELBO}(q(\theta)) = \int_\Theta \prod_{k=1}^{K} q(\theta_k) \log \pi(\theta, X) d\theta - \int_\Theta \prod_{k=1}^{K} q(\theta_k) \log q(\theta_k) d\theta$$

Consider maximising the ELBO by taking one θ_k at a time. Thus, by isolating the term $q(\theta_k)$,

$$\int q(\theta_k) \left[\int \log \pi(\theta, X) \prod_{h \neq k} q(\theta_h) d\theta_{(-k)} \right] d\theta_k - \int q(\theta_k) \log q(\theta_k) d\theta_k + C_k$$

If we introduce the quantity $\log \pi^*(\theta_k, X) = \mathbb{E}_{-k}(\log \pi(\theta, X)) + C$, and rearrange the terms, we get

$$\text{ELBO}(q(\theta)) = \int q(\theta_k) \log \frac{\pi^*(\theta_k, X)}{q(\theta_k} d\theta_k + C_k^* = -KL(q(\theta_j) || \pi * (\theta_k, X)) + C_k^*$$

Derivation of the CAVI algorithm [2]

The above expression implies that the local maximisation of ELBO($q(\theta)$) with respect to the k-th term of $q(\theta)$ is obtained by setting

$$\hat{q}(\theta_k) \propto \exp\left(\mathbb{E}_{-k} \log \pi(\theta, X)\right), \qquad \forall k = 1, \cdots, K$$

- In practice, the above expectation is often simple to compute, and specific kernels can usually be recognised, as in the Gibbs sampler.
- In the CAVI algorithm, we iteratively update the quantities $q(\theta_k)$ using the locally maximised terms given the others.
- By construction, CAVI produces a monotonic sequence converging to a local maximum of ELBO.

Comments

The CAVI is an appealing algorithm for maximising the ELBO under the mean-field assumption. However,

- In principle, one could use any other optimiser
- The necessary computations and expectations are usually easy to perform if the full conditional distributions belong to some exponential family.
- The algorithm stops whenever the ELBO sequence has converged.

Expectation Propagation

Expectation propagation (EP) is an iterative algorithm in which a target density $\pi(\theta|X)$ is approximated by a density $g(\theta)$ from some specified parametric family

Assume that the target density $\pi(\theta)vertX)$ has some convenient factorization up to proportion,

$$\pi(\theta|X) = \prod_{k=0}^{K} \pi_k(\theta|X)$$

In Bayesian inference, one can assign, for example, factor $k = 0$ to the prior $\pi(\theta)$ and factors 1 through K as the likelihood for the data partitioned into K parts $p(y_k|\theta)$ that are independent given the model parameters.

The algorithm works by iteratively approximating $\pi(\theta|X)$ with a density $g(\theta)$ which admits the same factorization

$$g(\theta) = \prod_{k=0}^{K} g_k(\theta)$$

and using some suitable initialization for all $g_k(\theta)$.

Expectation Propagation

At each iteration of the algorithm, and for $k = 0, \ldots, K$, we take the current approximating function $g(\theta)$ and replace $g_k(\theta)$ by the corresponding factor $\pi_k(\theta)$ from the target distribution. Accordingly, one defines the cavity distribution,

$$g_{-k}(\theta) \propto g(\theta)/g_k(\theta),$$

and the tilted distribution,

$$g_k^*(\theta) \propto \pi_k(\theta) g_{-k}(\theta)$$

The algorithm proceeds by first constructing an approximation $g_{new}(\theta)$ to the tilted distribution $g_k^*(\theta)$.

After this, an updated approximation to the target density $\pi_k(\theta|X)$ can be obtained as

$$g_k^{new}(\theta) \propto g^{new}(\theta)/g_{-k}(\theta),$$

and then iterate these updates.

The global approximation $g(\theta)$ and the site approximations $g_k(\theta)$ are restricted to be in a selected exponential family, such as the multivariate normal.