An extension of the Unified Skew-Normal family of distributions with application to Bayesian binary regression

Brunero Liseo

Sapienza Università di Roma

brunero.liseo@uniroma1.it

Joint work with Paolo Onorati

O'Bayes Conference, Santa Cruz (CA)

September 7^{th}, 2022

Outline

- We present a general Bayesian methodology for implementing binary regression models
- Our methods aims to
  - extend the approach described in [Durante(2019)] for the Probit model with a Gaussian Prior
  - provide a competitive alternative to existing methods [Polya-Gamma technique (Polson at al (2013)]; [Holmes and Held(2006)])

Ingredients:

- ◇ The Unified Skew Normal (SUN) class of densities
- ◇ Scale mixtures of Gaussian distributions
- ◇ Kolmogorov distribution
- ◇ Gibbs sampler

Prequel

The Unified Skew-Normal density has been introduced by [Arellano-Valle and Azzalini(2006)], but see also [O'Hagan and Leonard(1976)] for a proto-Bayesian use. Among several representations, it can be considered as a multivariate Gaussian with linear constraints.

$$Y = \xi + \mathrm{diag}^{1/2}(\Omega) Z \,|\, (U + \tau > 0)$$

with

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

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

The density function

It includes the computation of two CDFs of a multivariate Gaussian density

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

$\tau = 0 \Longrightarrow$ Skew-Normal family
$\Delta = 0$ or $m = 0 \Longrightarrow$ Normal family

A different representation

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

with $A \in \mathbb{R}^{d \times m}, b \in \mathbb{R}^m$.
This way, $T \perp\!\!\!\perp Z$ and $T \sim N_m(0, \Theta)$, with

$$\Theta = \mathrm{diag}^{-\frac{1}{2}}\left(\Gamma - \Delta'\bar{\Omega}^{-1}\Delta\right)\left(\Gamma - \Delta'\bar{\Omega}^{-1}\Delta\right)\mathrm{diag}^{-\frac{1}{2}}\left(\Gamma - \Delta'\bar{\Omega}^{-1}\Delta\right),$$

$$A = \mathrm{diag}^{-\frac{1}{2}}\left(\Gamma - \Delta'\bar{\Omega}^{-1}\Delta\right)\Delta'\bar{\Omega}^{-1}$$

and

$$b = \mathrm{diag}^{-\frac{1}{2}}\left(\Gamma - \Delta'\bar{\Omega}^{-1}\Delta\right)\tau.$$

SUN family and Probit model

[Durante(2019)] discovered a central role of the SUN density in Bayesian probit models.
Starting from a normal prior for the coefficients $\boldsymbol{\beta} \sim N_p(\xi, \Omega)$ the posterior for β after producing a probit likelihood, belongs to the SUN family

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

Remarks:

- The previous stochastic representation can be suitably used for posterior sampling
- The algorithm is particularly efficient in the $p > n$ case [Botev(2017)]

Extending the SUN family

We construct a larger class of densities, named perturbed SUN (pSUN) via the replacement of φ and Φ with scale mixtures of Gaussian densities.

This is done with the goal of finding a more general conjugacy in the Bayesian analysis of binary regression models.

Assume that $Z = \mathrm{diag}^{1/2}(W)R$ and $T = \mathrm{diag}^{1/2}(V)S$, with

$$\begin{aligned} V \sim Q_V(\cdot) \quad &\perp\!\!\!\perp \quad S \sim N_m(0, \Theta) \\ W \sim Q_W(\cdot) \quad &\perp\!\!\!\perp \quad R \sim N_d(0, \bar{\Omega}), \end{aligned}$$

The pSUN class is defined as the expression (1)

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

with the above assumptions on Z and T. Then,

$$pSUN_{d,m}(Q_V, \Theta, A, b, Q_W, \Omega, \xi).$$

The density of a pSUN

Let $Y \sim pSUN_{d,m}(Q_V, \Theta, A, b, Q_W, \Omega, \xi)$. Then

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

with

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

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

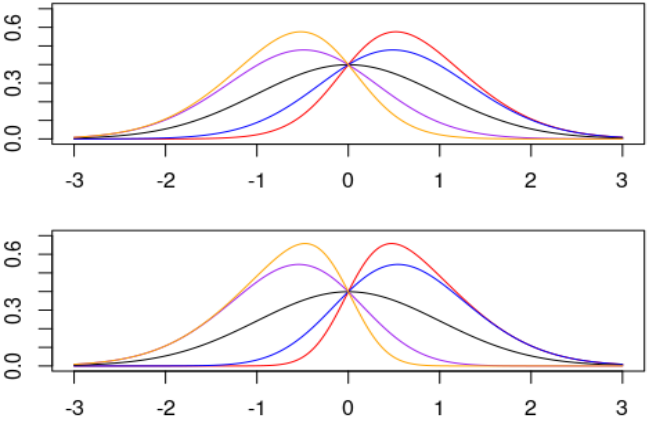
and

$$\Psi_{Q_V, \Theta, A, Q_W, \bar{\Omega}}(b) = \mathrm{P}(T - AZ \leq b)$$

$$T \sim \Phi_{\Theta, Q_V}(\cdot) \perp\!\!\!\perp Z \sim \Phi_{\bar{\Omega}, Q_W}(\cdot)$$

Some pSUN densities

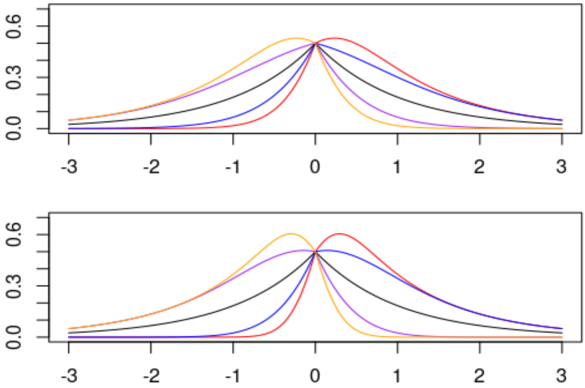
Logit: top: N(0,1); $V \sim LK(\cdot); W = 1, A = 3, b = 0$
$V \sim LK(\cdot); W = 1, A = 1.5, b = 0$



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

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

The MGF of a pSUN

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

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

with

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

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

$$f_{\widetilde{Z}_k}(x) = \frac{e^{k'x} f_Z(x)}{M_Z(k)}.$$

Sampling a pSUN

We adopted a Gibbs algorithm:

- Key aspect: one must be able to sample from the f.c.'s $W|Z$ and $V|T$.
- It is not always easy, and it depends on the specific values of $\Theta, \bar{\Omega}$, and the form of $Q_W(\cdot)$ and $Q_V(\cdot)$.
- Relatively simple in the most popular versions of the Bayesian binary regression.

Sampling a pSUN

At time t:

$$\text{Sample } V_{t+1} \sim V|T = T_t$$
$$\text{Sample } W_{t+1} \sim W|Z = Z_t$$

In order to sample $Z_{t+1}, T_{t+1} \sim Z, T|T \leq AZ + b, W_{t+1}, V_{t+1}$ do the following steps: set $\Theta_V = \mathrm{diag}^{1/2}(V)\Theta\,\mathrm{diag}^{1/2}(V)$ and $\bar{\Omega}_W = \mathrm{diag}^{1/2}(W)\bar{\Omega}\,\mathrm{diag}^{1/2}(W)$

$$\text{Set } \Sigma_\varepsilon = \Theta_{V_{t+1}} + A\bar{\Omega}_{W_{t+1}}A'$$
$$\text{Sample } \varepsilon \sim TN_m(-\infty, -b, 0, \Sigma_\varepsilon)$$
$$\text{Set } H_\mu = \bar{\Omega}_{W_{t+1}}A'\Sigma_\varepsilon^{-1}$$
$$\text{Set } H_\Sigma = (I - H_\mu A)\,\bar{\Omega}_{W_{t+1}}$$
$$\text{Sample } Z_{t+1} \sim N_d(H_\mu\varepsilon, H_\Sigma)$$
$$\text{Set } T_{t+1} = AZ_{t+1} - \varepsilon$$
$$\Longrightarrow Y_{t+1} = \xi + \mathrm{diag}^{1/2}(\Omega)Z_{t+1}$$

Linear Symmetric Binary Regression

Consider a general version of the model as

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

- $\Lambda : \mathbb{R} \to [0, 1]$ 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 $\Lambda_n(x) = \prod_{i=1}^n \Lambda(x_i)$ and $B_r = [2\,\mathrm{diag}(r) - I_n]$ for $r \in \{0, 1\}^n$.
The likelihood function of a LSBR is

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

Conjugacy for Linear Symmetric Binary Regression (LSBR)

Theorem

Consider a Bayesian LSBR model and assume

$$\beta \sim pSUN_{p,m}(Q_{V_0}, \Theta, A, b, Q_W, \xi, \Omega).$$

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

$$\beta | Y = y \sim pSUN_{p,m+n}\left(Q_{V_0} Q_{V^*}^n, \Theta^*, A^*, \begin{bmatrix} b \\ B_y X \xi \end{bmatrix}, Q_W, \xi, \Omega\right),$$

with

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

and $Q_{V_0} Q_{V^*}^n([x_1, x_2]') = Q_{V_0}(x_1) \prod_{i=1}^n Q_{V^*}(x_{2,i})$

Computation

- In order to produce a posterior sample with the Gibbs algorithm, one must be able to sample from the full conditional distributions of V and W.
- W: This is relatively simple when $\pi(\beta)$ either has an elliptical structure or it has independent components. For example, the SGH [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: It depends on the link function $\Lambda(\cdot)$. Simpler when Θ is diagonal; (independently sample $V_i | T_i \, i = 1, 2, \ldots, n + m$. This happens, for example, when $m = 0$ or $m = 1$.

Bayesian Logistic Regression

- The popular logistic regression model is a special case of those discussed in the previous Theorem
- The logistic distribution admits a representation in terms of a scale mixture of Gaussian distributions; see [Andrews and Mallows(1974)] and [Stefanski(1991)].

In fact,

$$T_i | V_{0,i} \sim N(0, 4V_{0,i}^2) \text{ and } V_{0,i} \sim K(\cdot) \Longrightarrow T_i \sim Logis(0,1)$$

that is

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

Kolmogorov's distribution

We will use the logistic Kolmogorov distribution:

$$V_i = 4V_{0,i}^2 \ , \ V_{0,i} \sim K(\cdot)$$

We denote it by $V_i \sim \mathrm{LK}(\cdot)$; the density is

$$\mathrm{lk}(v) = \begin{cases} v^{-\frac{5}{2}}\sqrt{2\pi}\sum_{j=1}^{+\infty}\left((2j-1)^2\pi^2 - v\right)\exp\left(-\frac{(2j-1)^2\pi^2}{2v}\right) & 0 < v \leq v^* \\ \sum_{j=1}^{+\infty}(-1)^{j-1}j^2\exp\left(-\frac{j^2 v}{2}\right) & v > v^* \end{cases}$$

for some $v^* > 0$; see [Onorati and Liseo(2022)] for details. For numerical reasons, we set $v^* \approx 1.98$ and truncate both series to the first 15 terms.

Comments

- [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 some improvements in terms of speed.
- We do: $V, W | T, Z$ and then $T, Z | V, W$
  [Holmes and Held(2006)]: $V, W | T, Z$; then $T - AZ | Z, V, W$
  and then $Z | T - AZ, V, W$
  where, in both cases, $\beta = \xi + \mathrm{diag}^{1/2}(\Omega) Z$.

Technical details

The hard step is "how to sample" from the f.c. of $V|T, \beta, W, Y = V|T$

- Notice that the first m components of $V|T$ are independent of the last n ones, and they only depend on the prior distribution.
- focus on the last n components of $V|T$: they are mutually independent so one only needs to sample from $V_i|T_i\,, i = m+1, m+2, \ldots, m+n$.
- we adopt an acceptance-rejection algorithm.

Simulation Study

Both in the probit and in the logit case:

- Priors: pSUN with weakly informative hyper-parameters in the spirit of Gelman et al. (2008) , i.e.

$$m = 0, \xi = 0_p, \Omega = \text{diagonal matrix}$$

$\Longrightarrow \pi(\beta)$ will be unimodal and symmetric about the origin.
Probit model implies $V_1 = V_2 = \cdots = V_n = 1$.
We consider 3 different priors

A. a Gaussian prior $(W_1 = W_2 = \cdots = W_p = 1)$ [Durante(2019)]
B. a multivariate Laplace with independent components $(W_1, W_2, \ldots W_p \overset{iid}{\sim} \text{Exp}\,(1/2))$
C. Dirichlet-Laplace prior [Bhattacharya et al. (2015], with a discrete uniform prior on the Dirichlet parameter, in $(0, 1]$ $\{1/300 \times j, j = 1, 2, \ldots, 300\}$.

Simulation Study: Ω values (Probit case)

The diagonal components of Ω were obtained, adapting a suggestion in Gelman et al.(2008)

$$\begin{aligned} \text{Gaussian:} \qquad & \omega_{11} = 100, \omega_{22} = \cdots = \omega_{pp} = 42.25 \\ \text{Laplace with indep. components:} \qquad & \omega_{11} = 100; \omega_{22} = \cdots = \omega_{pp} = 6.25. \end{aligned}$$

Simulation Study: Ω values (Logit case)

Logit model implies $V_1, V_2, \ldots V_n \stackrel{\texttt{iid}}{\sim} K(\cdot)$

Centred Normal:	$\omega_{11} = 256; \omega_{22} = \cdots = \omega_{pp} = 25;$
Laplace with indep. components:	$\omega_{11} = 210.25; \omega_{22} = \cdots = \omega_{pp} = 14.0625$

Simulation scheme: $p = 10$

---

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

- sample each covariate value indep $X_{ij}^{(g)} \sim N(0,1)$ and transform column of $X^{(g)}$ to have a s.d. $= 0.5$

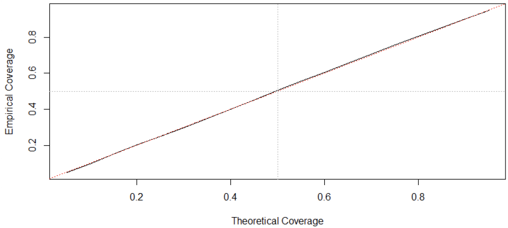
  for all model/prior combination
- if not DL, sample $\Sigma^{(g,h)} \sim W$ otherwise set $\Sigma_{g,h} = I$ and $\alpha \sim \pi(\alpha)$ sample $\beta \sim \pi_h(\beta | \Sigma_{g,h})$
- sample $Y_i^{(g,h)} \overset{ind}{\sim} Be(\Lambda_h(X_i'^{(g)} \beta_{True}^{(g,h)}))$
- draw N values from the posterior distribution of β
- compute the empirical quantiles of level $\gamma \in \{5/100 \times j, j = 1, 2, \ldots, 19\}$

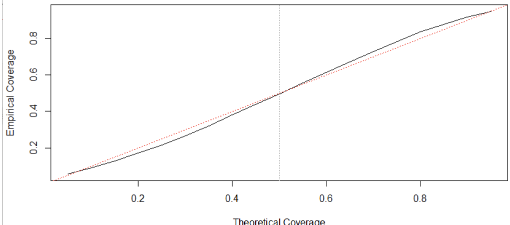
$\Longrightarrow$ evaluate the frequentist coverage comparing the quantiles with $\beta_{True}^{(g,h)}$

number of iteration in the Gibbs sampler: 10^4

Simulation Study: Ω values (Results)

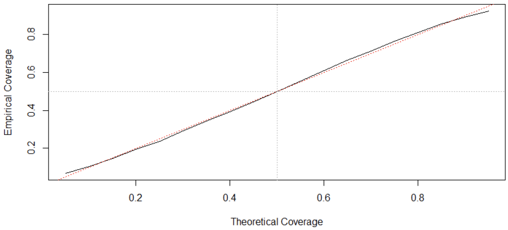
Logit model Frequentist coverage of priors in repeated sampling: Gaussian and Indep. Laplace

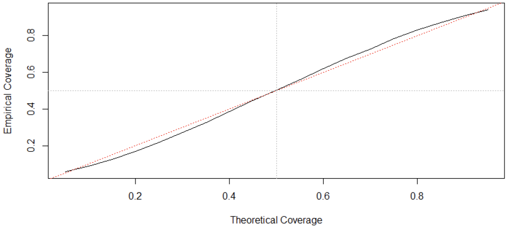




Simulation Study: Ω values (Results)

Probit model Frequentist coverage of priors in repeated sampling: Dirichlet-Laplace and Indep. Laplace

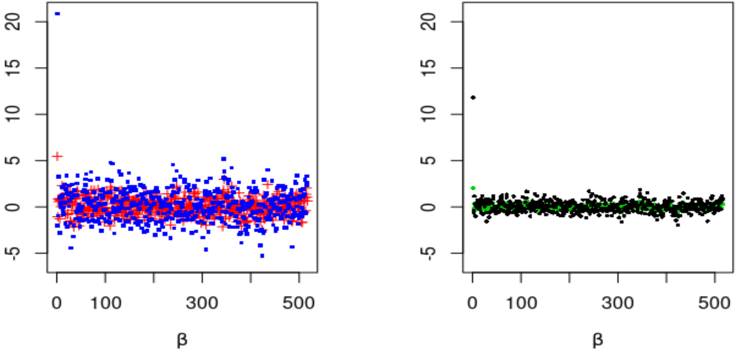




Cancer SAGE

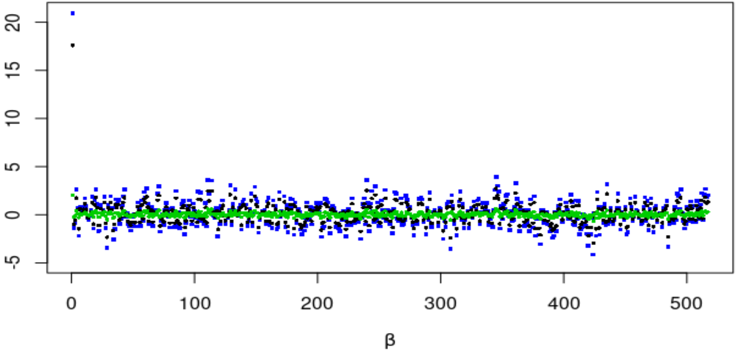
Discussed in [Durante(2019)]: a $p > n$ case:
$n = 74$ normal and cancerous biological tissues at 516 different tags.
Of interest: to quantify the effects of gene expressions on the probability of a cancerous tissue and predicting the status of new tissues as a function of the gene expression.
Gene expressions standardized with mean 0 and $\sigma = 0.5$.
When $p > n$ the prior input is decisive

Cancer SAGE



Probit model: Posterior means of the 516 β coefficients + intercept. Left: Durante's prior; Gaussian prior Right: Laplace with independent components (black), and Dirichlet-Laplace

Cancer SAGE



Logit model: Posterior means of the 516 β coefficients + intercept.

Gaussian prior; Laplace with independent components Dirichlet-Laplace .

Objective Bayes

The general expression of a pSUN prior for the β vector is

$$\beta \sim pSUN_{m,p}(Q_V, \Theta, A, b, Q_W, \xi, \Omega)$$

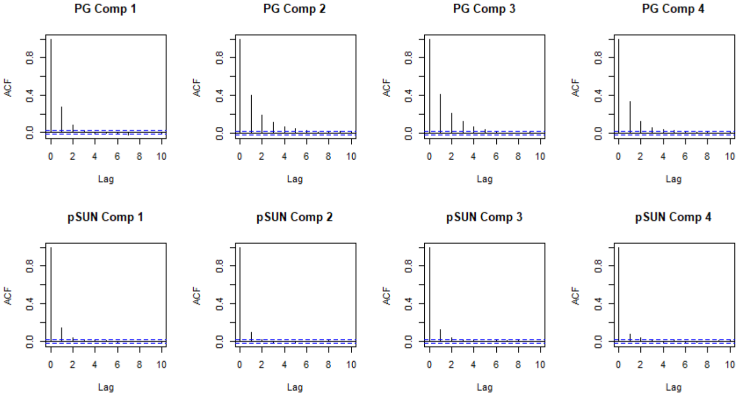
The natural *objective* version is then obtained by setting

m	Q_V	Θ	A	b	ξ
0	NA	NA	NA	NA	0

- Q_W and Ω are the only quantities to specify.
- For example, the adaptation of a sort of g-prior for binary responses (Marin & Robert, 2006) would correspond to $\Omega = (\boldsymbol{X'X})^{-1}$ and $W_1 = W_2 = \cdots = W_p = w$ and $\pi(w) \propto w^{-3/4}$.
- A weakly informative prior can be obtained by mimicking the approach described for the logit model in Gelman et al. (2008)

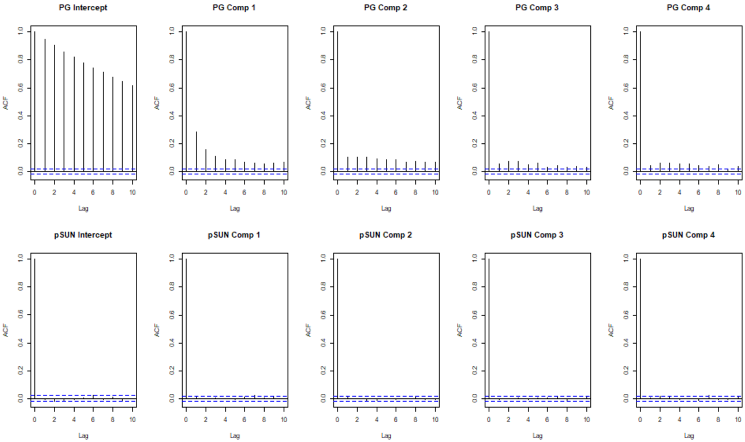
Comparison with Polson et al. (2013)

- with a small dataset ($n = 100, p = 4$):
  Polya-Gamma alg. takes 13 seconds with a $C++$ code. our algorithm is much slower [euphemism . . .] (5 minutes with a R code). However our ACF are much better



Comparison with Polson et al. (2013)

- with Cancer SAGE dataset ($n = 74, p = 517$): Polya-Gamma alg. is four time slower than pSUN (103 minutes vs 25 minutes), and *ACF* are still better



Future development

- Botev & L'Ecuyer (2015) have proposed an efficient method for simulating from a multivariate truncated Student t distribution. It works fine up to 100 dimensions
- This approach can be useful in our context for evaluating the normalizing constant of the posterior distribution. This can be suitably used for two different goals
  - providing an exact i.i.d. sampler
  - model selection via Bayes factor
- Make the algorithm faster in C++
- Semiparametric generalisations (see Paolo's poster)
- Tobit models

References I

D. F. Andrews and C. L. Mallows.
Scale mixtures of normal distributions.
J. Roy. Statist. Soc. Ser. B, 36:99–102, 1974.

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

O. Barndorff-Nielsen.
Exponentially decreasing distributions for the logarithm of particle size.
In *Proc. Royal Soc. Series A, Math. and Phys. Sci.* , , pages 401–409. The Royal Society, London, 1977.

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.
ISSN 1369-7412.
doi: 10.1111/rssb.12162.
URL `https://doi.org/10.1111/rssb.12162`.

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

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

References II

Anthony O'Hagan and Tom Leonard.
Bayes estimation subject to uncertainty about parameter constraints.
Biometrika, 63(1):201–203, 04 1976.

Paolo Onorati and Brunero Liseo.
Random Number Generator for the Kolmogorov Distribution.
arXiv2208.13598, 2022.
URL `https://arxiv.org/abs/2208.13598`.

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

Leonard A. Stefanski.
A normal scale mixture representation of the logistic distribution.
Statist. Probab. Lett., 11(1):69–70, 1991.