

Supporting Information for "The Scale Transformed Power Prior for Use with Historical Data from a Different Outcome Model"

Ethan M. Alt | Brady Nifong | Xinxin Chen | Matthew A. Psioda | Joseph G. Ibrahim

A | APPENDIX A: THE EXPRESSION FOR THE JACOBIAN IN THE STRAPP

In general, the transformation implied by

$$I_0^{1/2}(\boldsymbol{\eta})\boldsymbol{\eta} = I_1^{1/2}(\boldsymbol{\theta})\boldsymbol{\theta}, \quad (1)$$

cannot be calculated algebraically. However, the Jacobian matrix can be calculated via implicit derivation as shown below.

$$\begin{aligned} \left[\left\{ \frac{d}{d\boldsymbol{\eta}} I_0^{1/2}(\boldsymbol{\eta}) \right\} \boldsymbol{\eta} + I_0^{1/2}(\boldsymbol{\eta}) \right] \frac{d\boldsymbol{\eta}}{d\boldsymbol{\theta}} &= \left\{ \frac{d}{d\boldsymbol{\theta}} I_1^{1/2}(\boldsymbol{\theta}) \right\} \boldsymbol{\theta} + I_1^{1/2}(\boldsymbol{\theta}) \\ \Rightarrow \frac{d\boldsymbol{\eta}}{d\boldsymbol{\theta}} &= \left[\left\{ \frac{d}{d\boldsymbol{\eta}} I_0^{1/2}(\boldsymbol{\eta}) \right\} \boldsymbol{\eta} + I_0^{1/2}(\boldsymbol{\eta}) \right]^{-1} \left[\left\{ \frac{d}{d\boldsymbol{\theta}} I_1^{1/2}(\boldsymbol{\theta}) \right\} \boldsymbol{\theta} + I_1^{1/2}(\boldsymbol{\theta}) \right]. \end{aligned}$$

The matrix $\left\{ dI_1^{1/2}(\boldsymbol{\theta})/d\boldsymbol{\theta} \right\} \boldsymbol{\theta}$ can be written as $(\{dI_1^{1/2}(\boldsymbol{\theta})/d\theta_0\}\boldsymbol{\theta}, \dots, \{dI_1^{1/2}(\boldsymbol{\theta})/d\theta_{p-1}\}\boldsymbol{\theta})$. The derivative can be decomposed using a direct application of the product rule, for $j = 0, \dots, p-1$, as

$$dI_1(\boldsymbol{\theta})/d\theta_j = I_1^{1/2}(\boldsymbol{\theta})\{dI_1^{1/2}(\boldsymbol{\theta})/d\theta_j\} + \{dI_1^{1/2}(\boldsymbol{\theta})/d\theta_j\}I_1^{1/2}(\boldsymbol{\theta}). \quad (2)$$

Equation (2) can be expressed in the form of the Sylvester equation¹ which allows for vectorized representations of the needed derivatives². Let \mathbf{I}_p denote the $p \times p$ identity matrix. Then the required derivatives may be represented as

$$\text{vec} \left(\frac{dI_1^{1/2}(\boldsymbol{\theta})}{d\theta_j} \right) = \left\{ I_1^{1/2}(\boldsymbol{\theta}) \otimes \mathbf{I}_p + \mathbf{I}_p \otimes I_1^{1/2}(\boldsymbol{\theta}) \right\}^{-1} \text{vec} \left(\frac{dI_1(\boldsymbol{\theta})}{d\theta_j} \right), \quad (3)$$

where $\text{vec}(\cdot)$ denotes the vectorization of a matrix in which columns are stacked to convert a $n \times p$ matrix into a $np \times 1$ vector.

The derivative of $I_0^{1/2}(\boldsymbol{\eta})$ is calculated analogously.

B | APPENDIX B: THE LOCALLY ONE-TO-ONE TRANSFORMATION PROPERTY

To understand the implications of the local one-to-one property of the straPP transformation, we consider a single parameter logistic model with parameter η (success probability $\pi(\eta) = e^\eta / (1 + e^\eta)$). For a random sample of size n , it follows that the Fisher information has a simple closed form given by

$$I(\eta) = n\pi(\eta)(1 - \pi(\eta)),$$

where $\pi(\eta) = e^\eta / (1 + e^\eta)$. Thus,

$$I(\eta)^{1/2}\eta = \sqrt{n\pi(\eta)(1 - \pi(\eta))}\eta.$$

A plot of η by $I(\eta)^{1/2}\eta$ is given in Figure 1 for the case where, without loss of generality, we take $n = 1$ (equivalently, the average information in the non-iid setting). Thus, for $\eta \in [-2.4, 2.4]$, the function is one-to-one. This corresponds to probabilities $\pi(\eta) \in$

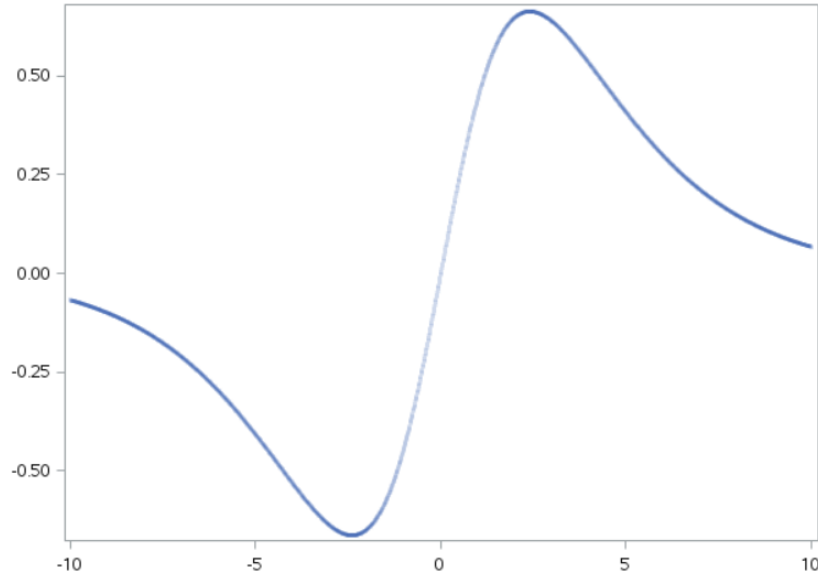


FIGURE 1 Plot of η (x-axis) by $I(\eta)^{1/2}\eta$ (y-axis).

$[0.084, 0.916]$. Thus, we only recommend considering the straPP transformation in such cases where response probabilities are not too extreme. Practitioners should choose initial priors that are informative enough to ensure the prior for η is consistent with the region where the straPP transformation is one-to-one.

C | APPENDIX C: PERFORMING MCMC WHEN USING A STRAPP

In many cases, one will not be able to solve for $\boldsymbol{\eta}$ in (1). When the Fisher information matrix for the historical data model does not depend on the regression parameters, such as when the historical data model is a linear model, one can solve algebraically for

$$\boldsymbol{\eta} = I_0^{-1/2} I_1^{1/2}(\boldsymbol{\theta}) \boldsymbol{\theta} = g(\boldsymbol{\theta})$$

and derive the straPP as a transformed power prior. In that case, MCMC proceeds in the obvious way. When the Fisher information matrix for the historical data depends on the regression parameters but the Fisher information matrix for the current data model does not, one can solve for

$$\boldsymbol{\theta} = I_1^{-1/2} I_0^{1/2}(\boldsymbol{\eta}) \boldsymbol{\eta} = g^{-1}(\boldsymbol{\eta})$$

and perform the analysis using *complementary sampling*. To understand this strategy, consider the (intractable) posterior distribution based on an analysis with the straPP, which is given by

$$\pi_s(\boldsymbol{\theta} \mid D_1, D_0) \propto \mathcal{L}(\boldsymbol{\theta} \mid D_1) \mathcal{L}(g(\boldsymbol{\theta}) \mid D_0)^{a_0} \pi_0(g(\boldsymbol{\theta})) \left| \frac{dg(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \right|,$$

where we now use D_1 to represent a current data set. Consider the transformation $\boldsymbol{\eta} = g(\boldsymbol{\theta})$. It is straightforward to show that applying the transformation leads to

$$\pi_s(\boldsymbol{\eta} \mid D_1, D_0) \propto \mathcal{L}(g^{-1}(\boldsymbol{\eta}) \mid D_1) \mathcal{L}(\boldsymbol{\eta} \mid D_0)^{a_0} \pi_0(\boldsymbol{\eta}), \quad (4)$$

which does not include a determinant term and is equivalent to parameterizing the current data model in terms of $\boldsymbol{\eta}$ and fitting the model using a power prior. Note that in this case, the function $g^{-1}(\boldsymbol{\eta})$ is readily available and so fitting this model with MCMC is again straightforward. Samples for $\boldsymbol{\eta}$ can be obtained from the posterior in (4) and transformed according to $\boldsymbol{\theta} = g^{-1}(\boldsymbol{\eta})$ in order to obtain the required samples for the parameters in the current data model.

The most challenging case occurs when the straPP equation cannot be solved for either parameter. In this case, MCMC using Hamiltonian Monte Carlo is straightforward but requires a non-linear equation solver. For example, a value of $\boldsymbol{\eta}$ might be proposed at a given MCMC iteration and then, using a non-linear equation solver, the algorithm must compute the corresponding proposed value of $\boldsymbol{\theta}$ in order to evaluate the posterior kernel to construct the rejection ratio. Such a feature is readily available in `rstan`³ but can require substantial computation time.

D | APPENDIX D: THE STRAPP FOR MULTIPLE HISTORICAL DATA SETS

Similar to the power prior, the straPP can be extended for use with multiple historical data sets. Suppose there are K historical data sets available, denoted D_{0k} for $k = 1, \dots, K$, all with the same outcome distribution. Let $\mathbf{D}_0 = (D_{01}, \dots, D_{0K})$ and $\mathbf{a}_0 =$

(a_{01}, \dots, a_{0K}) , where a_{0k} is the weight associated with the k^{th} historical data set. Then, the straPP for multiple historical data sets can be written as

$$\pi_{ms}(\theta \mid \mathbf{D}_0) \propto \left\{ \prod_{k=1}^K \mathcal{L}(g(\theta) \mid D_{0k})^{a_{0k}} \pi_0(g(\theta)) \right\} \left| \frac{dg(\theta)}{d\theta} \right|,$$

where $g(\theta)$ is the function induced by the transformation $I_0^{1/2}(\boldsymbol{\eta})\boldsymbol{\eta} = I_1^{1/2}(\theta)\theta$ and the covariate matrix used to formulate $I_0(\boldsymbol{\eta})$ and $I_1(\theta)$ is the stacked covariate matrix obtained by vertical concatenation of the covariate matrices from the K data sets.

E | APPENDIX E: EXPRESSION FOR THE JACOBIAN FOR GENERALIZED LINEAR MODELS WITH THE CANONICAL LINK

In this section, we assume that outcomes for the historical and current data arise from the class of generalized linear models (GLMs), as described in Section 4.4 of the paper. Further, we specify the canonical link for the historical and current data models. Then the Fisher information matrix from the historical data model for the regression parameter can be written as $I_0(\boldsymbol{\beta}_0) = \phi_0 X_0^T V_0(\boldsymbol{\beta}_0) X_0$ and the Fisher information matrix from the current data model can be written as $I_1(\boldsymbol{\beta}_1 \mid X_0) = \phi_1 X_0^T V_1(\boldsymbol{\beta}_1) X_0$, where for $k = 0, 1$, $V_k(\boldsymbol{\beta}_k) = \text{diag}(v_{ki}(\boldsymbol{\beta}_k))$, as stated in Section 4.4 of the paper. Then following Section 4.1 of the paper, we can write the derivatives as

$$\begin{aligned} \left\{ \frac{d}{d\boldsymbol{\beta}_0} I_0^{1/2}(\boldsymbol{\beta}_0) \right\} \boldsymbol{\beta}_0 &= \left(\left\{ \frac{d}{d\beta_{0,0}} I_0^{1/2}(\boldsymbol{\beta}_0) \right\} \boldsymbol{\beta}_0, \dots, \left\{ \frac{d}{d\beta_{0,p-1}} I_0^{1/2}(\boldsymbol{\beta}_0) \right\} \boldsymbol{\beta}_0 \right), \\ \left\{ \frac{d}{d\boldsymbol{\beta}_1} I_1^{1/2}(\boldsymbol{\beta}_1 \mid X_0) \right\} \boldsymbol{\beta}_1 &= \left(\left\{ \frac{d}{d\beta_{1,0}} I_1^{1/2}(\boldsymbol{\beta}_1 \mid X_0) \right\} \boldsymbol{\beta}_1, \dots, \left\{ \frac{d}{d\beta_{1,p-1}} I_1^{1/2}(\boldsymbol{\beta}_1 \mid X_0) \right\} \boldsymbol{\beta}_1 \right). \end{aligned}$$

Using the results from Section 4.1, for a given $j = 0, \dots, p-1$, we find

$$\begin{aligned} \text{vec} \left(\frac{dI_0^{1/2}(\boldsymbol{\beta}_0 \mid X_0)}{d\beta_{0,j}} \right) &= \left\{ I_0^{1/2}(\boldsymbol{\beta}_0) \otimes \mathbf{I}_{p \times p} + \mathbf{I}_{p \times p} \otimes I_0^{1/2}(\boldsymbol{\beta}_0 \mid X_0) \right\}^{-1} \\ &\quad \times \text{vec} \left(X_0^T \text{diag} \left(\frac{dv_{0i}(\boldsymbol{\beta}_0)}{d\beta_{0,j}} \right) X_0 \right), \\ \text{vec} \left(\frac{dI_1^{1/2}(\boldsymbol{\beta}_1 \mid X_0)}{d\beta_{1,j}} \right) &= \left\{ I_1^{1/2}(\boldsymbol{\beta}_1 \mid X_0) \otimes \mathbf{I}_{p \times p} + \mathbf{I}_{p \times p} \otimes I_1^{1/2}(\boldsymbol{\beta}_1 \mid X_0) \right\}^{-1} \\ &\quad \times \text{vec} \left(X_0^T \text{diag} \left(\frac{dv_{1i}(\boldsymbol{\beta}_1)}{d\beta_{1,j}} \right) X_0 \right). \end{aligned}$$

F | APPENDIX F: PROOF OF THEOREM 1

Let $\mathbf{Y}_0 \sim N_{n_0}(X_0 g(\boldsymbol{\beta}_1), \sigma_0^2 \mathbf{I}_{n_0})$ be the $n_0 \times 1$ vector of responses for the historical data and $\mathbf{Y}_1 \sim N_{n_1}(X_1 \boldsymbol{\beta}_1, \sigma_1^2 \mathbf{I}_{n_1})$ be the $n_1 \times 1$ vector of responses for the current data, with known historical and current data variance. Then, the posterior distribution

associated with an analysis based on the straPP and power prior are given by (5) and (6), respectively.

$$\beta_1 | Y_1, Y_0 \stackrel{\text{straPP}}{\sim} N_p(\hat{\beta}_s, \Sigma_s) \quad (5)$$

$$\beta_1 | Y_1, Y_0 \stackrel{\text{PP}}{\sim} N_p(\hat{\beta}_p, \Sigma_p), \quad (6)$$

where $\hat{\beta}_s = \Sigma_s \{(1/\sigma_1^2)X_1^T Y_1 + a_0/(\sigma_0\sigma_1)X_0^T Y_0\}$, $\Sigma_s = \sigma_1^2(X_1^T X_1 + a_0X_0^T X_0)^{-1}$, $\hat{\beta}_p = \Sigma_p \{(1/\sigma_1^2)X_1^T Y_1 + (a_0/\sigma_0^2)X_0^T Y_0\}$, and $\Sigma_p = \{(1/\sigma_1^2)X_1^T X_1 + (a_0/\sigma_0^2)X_0^T X_0\}^{-1}$.

Now let $\hat{\beta}_{s,1j}$ and $\hat{\beta}_{p,1j}$ denote the straPP and power prior estimator for β_{1j} , respectively, where β_{1j} is the $(j + 1)$ th element of β_1 ($j = 0, \dots, p-1$). We wish to find the smallest β_{1j} such that $\text{MSE}(\hat{\beta}_{s,1j}) \leq \text{MSE}(\hat{\beta}_{p,1j})$. Note that the MSE can be decomposed into the sum of the point estimator's variance and squared bias (i.e., $\text{MSE} = \text{Var} + [\text{Bias}]^2$).

First, we show that $\text{Bias}(\hat{\beta}_s) = \mathbf{0}$, and thus $\text{Bias}(\hat{\beta}_{s,1j}) = 0$. For the normal-normal case, the assumed relationship between the historical and current data model parameters is $\beta_0 = g(\beta_1) = (\sigma_0/\sigma_1)\beta_1$. Then $E(Y_0)/(\sigma_0\sigma_1) = X_0\beta_0/(\sigma_0\sigma_1) = X_0\beta_1/\sigma_1^2$. We can calculate the bias of the straPP estimator as

$$\begin{aligned} \text{Bias}(\hat{\beta}_s) &= E(\hat{\beta}_s) - \beta_1 = \Sigma_s \left\{ \frac{1}{\sigma_1^2} X_1^T E(Y_1) + \frac{a_0}{\sigma_0\sigma_1} X_0^T E(Y_0) \right\} - \beta_1 \\ &= \Sigma_s \left\{ \frac{1}{\sigma_1^2} (X_1^T X_1 + a_0 X_0^T X_0) \right\} \beta_1 - \beta_1 \\ &= \Sigma_s (\Sigma_s)^{-1} \beta_1 - \beta_1 = \mathbf{0}. \end{aligned}$$

For non-zero β_{1j} , it follows that

$$\begin{aligned} \text{Var}(\hat{\beta}_{s,1j}) \leq \text{Var}(\hat{\beta}_{p,1j}) + \left\{ \text{Bias}(\hat{\beta}_{p,1j}) \right\}^2 &\Leftrightarrow \text{Var}(\hat{\beta}_{s,1j}) \leq \text{Var}(\hat{\beta}_{p,1j}) + \beta_{1j}^2 \left\{ \text{Percent Bias}(\hat{\beta}_{p,1j}) \right\}^2 \\ &\Leftrightarrow \frac{\text{Var}(\hat{\beta}_{s,1j}) - \text{Var}(\hat{\beta}_{p,1j})}{\left\{ \text{Percent Bias}(\hat{\beta}_{p,1j}) \right\}^2} \leq \beta_{1j}^2. \end{aligned}$$

G | APPENDIX G: BINARY-NORMAL CASE – STRAPP TRANSFORMATION VIOLATED

We consider the binary-normal case where the parameters in the historical and current data models do not satisfy the assumption of the straPP transformation. The purpose of these simulations is to explore the robustness of the Gen-straPP to account for such violations. To operationalize this investigation, we assumed

$$\beta_1 = I_1^{-1/2} I_0^{1/2} (\beta_0) \beta_0 - I_1^{-1/2} c_0, \quad (7)$$

where $c_0^T = (0, c_{01})$ and c_{01} was varied. We considered the following inputs: $n_0 = 100$, $n_1 = 100$, $a_0 = 0.5$, $\beta_{00} = -0.5$, $\beta_{01} = 0.25$, $\sigma_1 = 2$, and $c_{01} \in \{-1.5, -1.4, \dots, 1.4, 1.5\}$. The values of the current data model parameters were then identified

by solving (7). Figure 1 panels (a)-(d) present results comparing performance characteristics of the straPP, the Gen-straPP, and the commensurate prior.

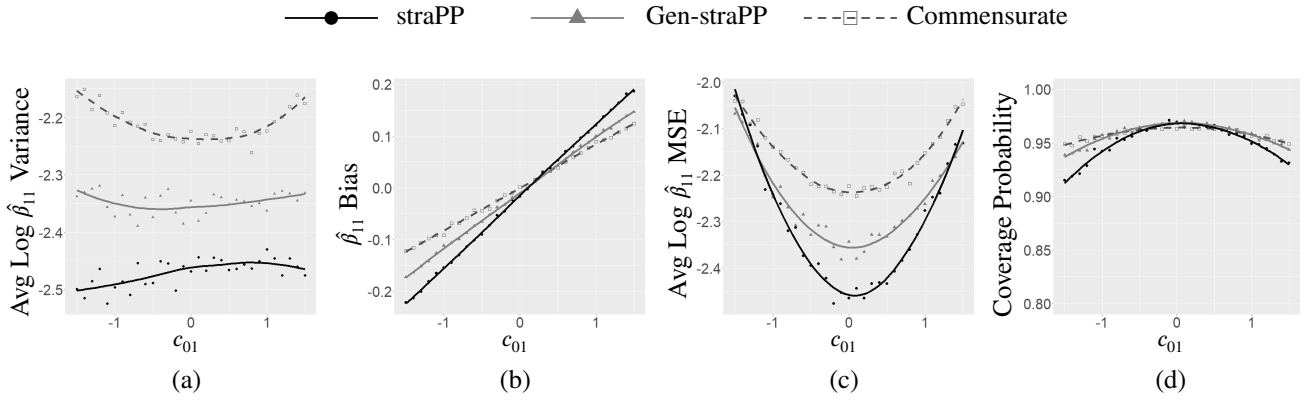


FIGURE 1 Panels (a)-(d) present the average log variance, bias, log MSE, and coverage probability for the posterior mean of β_{11} , respectively, as a function of c_{01} plotted on the x-axis for the straPP, Gen-straPP, and commensurate prior. straPP, scale transformed power prior.

The average log variance of the posterior mean based on the straPP is smaller than the Gen-straPP and commensurate prior, as seen in Figure 1(a). Figure 1(b) illustrates that the magnitude of the posterior mean estimator bias increases as c_{01} increases and is greater for the straPP compared to the Gen-straPP and commensurate priors. This illustrates the robustification provided by the Gen-straPP in terms of bias reduction when the assumption of the straPP does not hold. In Figure 1(c), the MSE for posterior mean based on the straPP is lower when the value of c_{01} is near zero but exceeds the MSE for the posterior mean based on the Gen-straPP when the quantity becomes sufficiently large in absolute value. The coverage probabilities based on an analysis with the Gen-straPP and commensurate prior are near or above 95% with the straPP-based coverage probabilities dipping to near 90% in the extreme cases.

H | APPENDIX H: BINARY-NORMAL SIMULATION SETUP

We generated $B = 5,000$ data sets. For each sampler, we used a burn-in of 5,000 samples and took 25,000 samples without thinning. All samplers were coded in the Stan programming language. The sample sizes were $n = 100$ and $n_0 = 100$.

For the historical data set, we generated $y_{0i} \sim \text{Ber}(\mu_{0i})$, where

$$\log\left(\frac{\mu_{0i}}{1 - \mu_{0i}}\right) = -0.5 + \beta_{01}x_{0i},$$

where $\beta_{01} \in \{0, 0.1, 0.2, \dots, 2.0\}$, and x_{0i} is a binary covariate.

The regression coefficients were obtained by solving the straPP transformation, i.e.,

$$\beta_1 = (\mathbf{X}'_0 \mathbf{X}_0)^{-1/2} [\mathbf{X}'_0 \mathbf{W}_0(\beta_0) \mathbf{X}_0]^{1/2} \beta_0.$$

For the current data set, we generated $y_i \sim N(\mu_{1i}, \sigma_1^2)$, where $\mu_{1i} = \beta_{10} + \beta_{11}x_i$, where x_i is a binary covariate, and $\sigma_1 = 2$.

The initial prior for the historical data was

$$\pi_0(\beta_0) = \phi(\beta_{00}|0, 1.645^2) \times \phi(\beta_{01}|0, 10^2)$$

The prior for the precision parameter $\kappa = \sigma^{-2}$ was

$$\pi(\kappa) = f_\Gamma(\kappa|0.001, 0.001),$$

where $f_\Gamma(\cdot|a, b)$ is the gamma density function with shape a and rate (inverse scale) parameter b .

For the Gen-StraPP, we elicited

$$\begin{aligned} c_0 | \sigma_{c_0}^2 &\sim N(0, \sigma_{c_0}^2), \\ \sigma_{c_0}^2 &\sim N^+(0, 1), \end{aligned}$$

where $N^+(\mu, \sigma^2)$ is a normal distribution with mean 0 and variance σ^2 truncated from below at zero.

For the Commensurate Prior, we elicited

$$\tau^{-1/2} \sim N^+(0, 1).$$

For the normalized version of the power prior and straPP, we elicited

$$\pi(\beta|D_0) \propto N(\hat{\beta}_0, I(\hat{\beta}_0)^{-1}/a_0),$$

with

$$\pi(a_0) \propto 1\{0.005 \leq a_0 \leq 1\},$$

for which the lower bound was chosen to facilitate computation.

I | APPENDIX I: POSTERIOR ESTIMATES FOR INTERCEPTS AND VARIANCE

TABLE 1 Posterior Estimates for Intercepts and Current Linear Regression Variance

Model	a_0	DIC	Historical Intercept		Current Intercept		Current Variance	
			Mean (SD)	95%HPD	Mean (SD)	95%HPD	Mean (SD)	95%HPD
Gen-straPP	0.10	2815.38	-1.12 (0.57)	(-2.35, -0.12)	47.09 (1.01)	(45.11, 49.07)	86.40 (6.26)	(75.17, 99.58)
straPP	0.25	2815.37	-1.11 (0.38)	(-1.93, -0.41)	46.93 (0.95)	(45.09, 48.81)	86.98 (6.34)	(75.44, 100.26)
RP	–	2816.65	—————	—————	47.07 (1.09)	(44.96, 49.24)	86.78 (6.39)	(75.13, 100.24)
PP	0.10	2816.44	-1.40 (0.89)	(-3.24, 0.27)	46.66 (0.87)	(44.97, 48.39)	87.07 (6.40)	(75.48, 100.54)
COM	–	2818.47	-1.38 (0.34)	(-2.07, -0.74)	46.29 (0.89)	(44.59, 48.08)	87.57 (6.48)	(75.81, 100.92)

Gen-straPP, generalized scale transformed power prior; straPP, scale transformed power prior; RP, reference prior; PP, power prior; COM, commensurate prior.

References

1. Sylvester J. Sur l'equations en matrices $px = xq$. *C.R. Acad. Sci. Paris* 1884; 99: 67-71, 115-116.
2. Laub AJ. *Matrix Analysis for Scientists and Engineers*. Society for Industrial and Applied Mathematics . 2005.
3. Stan Development Team . RStan: the R interface to Stan. 2022. R package version 2.21.5.