UNIVERSITY OF THE
FREE STATE
UNIVERSITEIT VAN DIE
VRYSTAAT
YUNIVESITHI YA
FREISTATA

UFS·UV
NATURAL AND
AGRICULTURAL SCIENCES
NATUUR- EN
LANDBOUWETENSKAPPE

MATHEMATICAL STATISTICS
AND ACTUARIAL SCIENCE
WISKUNDIGE STATISTIEK
EN AKTUARIËLE WETENSKAP

Bayesian Estimation of a Robust Alternative to the Probit Model for Classification and Prediction

M. J. von Maltitz and A. J. van der Merwe

February 9, 2015

This paper introduces a new robust Bayesian procedure for modelling ordinal categorical response data as a function of exogenous covariates. The modelling procedure expands on the existing literature by assuming that the (ordinal) categorical responses are linked to skew t-distributed latent data. The model's prediction/classification performance is compared with the existing Bayesian probit model in a simulation study covering various forms of latent data. The srobit model is shown to be marginally better than the probit under all data scenarios, at the expense of additional computational complexity.

Keywords: Bayesian estimation, Gibbs sampler, ordinal data, categorical regression, probit model, strobit model.

1. Introduction

A categorical response model is a regression model in which the dependent variable can take on one of a set of values. The probit model, one type of binary response model, assumes that there is an underlying, latent variable (not observed), which indicates in which category each observation belongs. This underlying variable can be a function of the observed covariates, with a Normally distributed error. There are other models that can map $(-\infty, \infty)$ data to the $(0, 1)$ space, for example the logit link function, and the complementary log-log. This paper builds on the Bayesian estimation processes of probit, set out by Albert & Chib (1993), the logit, set out by Groenewald & Mokgatlhe (2005), and the robit, set out by Liu (2005), and introduces a more robust model for the underlying latent variable, namely a model based on a skew adaptation of Student's t-distribution.

Since the model is based on a (robust) skew t-distribution, it will henceforth be referred to as the strobit model. This study is concerned with estimation of the strobit model for both binary and ordinal categorical responses, the latter being an extension of the former.

The Bayesian estimation procedure does not actually model a categorical response variable as a function of the predictors. Rather, it models the latent variable as a function of the predictors. This implies that the estimated regression parameters have no meaning, except for classification and prediction purposes. For this study, however, this is of no concern, because the model can be used for the prediction of a category for a new observation (or of an observation with missing response category). Thus, the estimation method of this regression model is suitable for sequential regression multiple imputation (SRMI)[1], for example. For more details on SRMI, see Raghunathan, Lepkowski, van Hoewyk & Solenberger (2001). The goal of the study is to determine whether or not a more robust model for the underlying latent variable leads to better classification of new observations (observations with missing binary or ordinal responses) when the underlying latent variable is misspecified.

This paper first reviews estimation procedures of the Bayesian probit model for binary and ordinal responses, as constructed by Albert & Chib (1993). We will then introduce methodology to estimate the parameters of a skew t-distribution, and incorporate this process into the estimation of the latent variable in the binary and ordinal response strobit models. Some practicalities will be discussed, after which the strobit will be tested on simulated data. Conclusions will be drawn based on the comparison between the probit and strobit models after these models are applied to categorical data that is built on both latent Normal and non-Normal assumptions.

2. Bayesian Estimation of the Probit Model

In this section, we review probit and ordered probit estimation as laid out by Albert & Chib (1993). It is important to understand the MCMC simulation procedure for this method, since it will be adapted for estimation of the strobit and ordered strobit models. Additionally, the probit and ordered probit models are compared to the strobit and ordered strobit models, respectively, in the simulation study.

2.1. Two-category probit model

Consider a binary outcome vector Y, and covariate matrix X with rows $x_1, \ldots, x_n$. Introduce n latent variables (one for each observation), $W_1, \ldots, W_n$, where the W_i are independent $N(x_i'\beta, 1)$, and define $Y_i = 2$ if $W_i > 0$ and $Y_i = 1$ otherwise. It can be shown that the Y_i are independent Bernoulli r.v. with $p_i = P(Y_i = 2) = \Phi(x_i'\beta)$. So the joint posterior of the unobservables is:

$$\pi(\beta, W|y) \propto \pi(\beta) \prod_{i=1}^{n} \left(I_{W_i>0} I_{y_i=2} + I_{W_i\leq 0} I_{y_i=1}\right) \phi\left(W_i; x_i'\beta, 1\right),$$

where the vector y represents the observed categorical data, $\pi(\beta)$ is the prior on β, I is an indicator function that takes the value 1 on the subscripted condition, and 0 otherwise, and $\phi(W_i; x_i'\beta, 1)$ is the Normal density function for the variable W_i with mean $x_i'\beta$, and variance 1.

[1] Also known as the fully conditional specification (FCS) approach to multiple imputation, or multiple imputation through chained equations (MICE).

The conditional posterior distributions (using diffuse priors) are as follows:

$$\beta|y, W \sim N\left(\left(X'X\right)^{-1}\left(X'W\right), \left(X'X\right)^{-1}\right) \tag{1}$$

$$\begin{aligned} W_i|y, \beta &\sim N\left(x_i'\beta, 1\right) \text{ truncated at the left by 0 if } y_i = 2 \\ W_i|y, \beta &\sim N\left(x_i'\beta, 1\right) \text{ truncated at the right by 0 if } y_i = 1 \end{aligned} \tag{2}$$

Thus for a Gibbs sampler to simulate draws from the joint posterior is given by the following sequential procedure:

1. Initialise $\beta^{(0)}$ using the least squares estimate $\left(X'X\right)^{-1}\left(X'y\right)$.
2. Generate a vector W from Equation (2), given the preceding draw of β.
3. Generate a new vector β from Equation (1), given the preceding draw of W.
4. Repeat steps 2 and 3 until convergence of W and β.

2.2. Ordinal probit model

Albert and Chib (1993) also described an approach for Bayesian estimation of an ordered probit, similar to the two-category estimation procedure. The first category split (between categories 1 and 2), γ_1, is pinned down on the latent variable at 0, as before. The second split, γ_2 (to differentiate between categories 2 and 3), becomes an additional parameter to be estimated in the Gibbs sampler. Similarly, if there are more than three categories, each additional boundary, γ_j, on the underlying latent variable is another parameter to estimate within the Gibbs sampler.

In the case of the ordered probit, given that γ is a vector of the J category boundaries on the latent variable, the joint posterior of the unobservables is (with diffuse priors):

$$\pi\left(\beta, \gamma, W|y\right) \propto \prod_{i=1}^{n}\left\{\left[\sum_{j=1}^{J} I_{Y_i=j} I_{\gamma_{j-1}<W_i<\gamma_j}\right]\phi\left(W_i; x_i'\beta, 1\right)\right\}$$

The conditional distributions for the $\gamma_j|W, Y$ are then:

$$U\left\{\max\left[\max\left(W_i : Y_i = j\right), \gamma_{j-1}\right], \min\left[\min\left(W_i : Y_i = j+1\right), \gamma_{j+1}\right]\right\} \tag{3}$$

In the Gibbs sampler, the $\gamma_j|W, Y$ parameters are drawn before the W_i and the parameter estimates. Thus, for a Gibbs sampler to simulate draws from the joint posterior proceed as follows:

1. Initialise β using the least squares estimate $\left(X'X\right)^{-1}\left(X'y\right)$.
2. Generate category splits from Equation (3), with γ_1 fixed at a latent value of 0, given the previously generated W.
3. Generate a new vector W from Equation (2), given the preceding draw of W and the draws of the γ_j.
4. Generate a new vector β from Equation (1), given the preceding draw of W.
5. Repeat steps 2–4 until convergence of W and the parameters.

3. The Skew Student t-Distribution

We follow the setup presented in Fonseca, Ferreira & Migon (2008, p. 326). Consider a linear regression model in which an observation vector $y = (y_1, \ldots, y_n)'$ satisfies

$$y = X\beta + Z\delta + \epsilon$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are the regression coefficients, δ is a skewness parameter, Z is a vector with elements $z_i > 0$, $i = 1, 2, \ldots, n$ as skewness coefficients, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ is the error vector and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. according to the Student-t distribution with location zero, scale parameter σ and ν degrees of freedom. Here $X = [x_1, \ldots, x_n]'$ is the $n \times p$ matrix of explanatory variables and is taken to be full rank p. We denote the model parameters by $\theta = (\beta, \delta, \sigma, \nu) \in \mathbb{R}^{p+1} \times (0, \infty)^2$. The likelihood function is given by:

$$L(\beta, \sigma, \nu | y, X) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)^n \nu^{n\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)^n \pi^{n/2} \sigma^n} \prod_{i=1}^{n} \left[\nu + \left(\frac{y_i - x_i'\beta - \delta z_i}{\sigma}\right)^2\right]^{-(\nu+1)/2}. \tag{4}$$

The likelihood for the t-distribution given in Equation (4) can be restructured as follows:

$$L \propto \prod_{i=1}^{n} \left(\frac{\lambda_i \tau}{2\pi}\right)^{\frac{1}{2}} \exp\left[-\frac{\tau}{2}\left(y_i - x_i'\beta - \delta z_i\right)^2\right] \times \prod_{i=1}^{n} \left[\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \lambda_i^{\nu/2-1} \exp\left(\frac{-\nu\lambda_i}{2}\right)\right] \tag{5}$$

where $\tau = \sigma^{-2}$ and the λ_i are weights indicating the influence of each observation on ν. Integrating out the λ_i in Equation (5) yields Equation (4).

3.1. Fitting the skew t-distribution

When the t-distribution is used for errors on the posterior predictive distribution, generating the imputations is simply a matter of applying the posterior-drawn regression parameters to the covariates and adding an appropriate t error. The challenge is to find the degrees of freedom for this error. This involves a Gibbs sampling process for the parameters β, τ, $z_i, i = 1, \ldots, n$, δ, $\lambda_i, i = 1, \ldots, n$, and ν, while ν itself is drawn via a Metropolis-Hastings algorithm in each step of the Gibbs sampler. The Gibbs sampler requires the formulation of the conditional posterior distributions for each of the parameters of the model.

For each observation $i, i = 1, \ldots, n$, and covariate $q, q = 0, 1 \ldots, p$, $\tilde{y}_{iq} = y_i - \beta_{-q} X_{-q} - \delta z_i$, with $-q$ representing all variables in X besides variable q. In other words, for $q = 0$:

$$\tilde{y}_{i0} = y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip} - \delta z_i$$

For $q = 1$:

$$\tilde{y}_{i1} = y_i - \beta_0 - \beta_2 x_{i2} - \beta_3 x_{i3} - \ldots - \beta_p x_{ip} - \delta z_i$$

For $q = 2, \ldots, p$:

$$\tilde{y}_{iq} = y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_{q-1} x_{i(q-1)} - \beta_{q+1} x_{i(q+1)} - \ldots - \beta_p x_{ip} - \delta z_i$$

Finally, for $q = p$:

$$\tilde{y}_{ip} = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_{p-1} x_{i(p-1)} - \delta z_i$$

We also define $\tilde{\tilde{y}}_i = y_i - \beta x_i - \delta z_i$ separate to $\hat{y}_i = y_i - \beta x_i$, where x_i is the i^{th} row of the data matrix, corresponding to the covariates for observation i.

With skewness a part of the $\tilde{y}_{iq}$, the same conditional distributions exist for the β_q:

$$\beta_q|y, \beta_{-q}, \tau, \Lambda \sim$$
$$N\left\{\left(\tau\sum_{i=1}^{n}\lambda_i x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2}\right)^{-1}\left(\tau\sum_{i=1}^{n}\lambda_i x_{iq}\tilde{y}_{iq} + \frac{\mu_{\beta_q}}{\sigma_{\beta_q}^2}\right), \left(\tau\sum_{i=1}^{n}\lambda_i x_{iq}^2 + \frac{1}{\sigma_{\beta_q}^2}\right)^{-1}\right\}, \tag{6}$$

where x_{iq} is element (i, q) of the data matrix X (and when $q = 0$, $x_{i0} = 1$ for all i), and μ_{β_q} and $\sigma_{\beta_q}^2$ are the conjugate Normal prior mean and variance for β_q respectively. Once again, $\mu_{\beta_q} = 0$ and $\sigma_{\beta_q}^2 = 10000$.

For τ, we have that:

$$\tau|y, \beta, \Lambda \sim \Gamma\left\{\frac{n}{2} + a_\tau, \left(\frac{1}{2}\sum_{i=1}^{n}\lambda_i\tilde{\tilde{y}}_i^2 + 2b_\tau\right)^{-1}\right\}, \tag{7}$$

where a_τ and b_τ are the conjugate Gamma prior parameters for τ, and Λ is the diagonal matrix with diagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_n$.. However, for the case of the strobit and ordered strobit models, without loss of generality, τ is fixed at 1, just as σ is fixed at 1 in the formulation of the probit estimation of Albert & Chib (1993).

The conditional posterior for the $z_i, i = 1, \ldots, n$ is derived to be:

$$z_i|y, \beta, \tau, \delta, \Lambda \sim N\left\{\left(\tau\lambda_i\delta^2 + 1\right)^{-1}\tau\lambda_i\delta\hat{y}_i, \left(\tau\lambda_i\delta^2 + 1\right)^{-1}\right\} I_{Z_i>0}, \tag{8}$$

where $I_{Z_i>0}$ is an indicator function to ensure that only positive z_i exist (in order to make sense of the sign of the skewness parameter δ).

The conditional posterior distribution of the skewness parameter, δ, is given can be shown to be:

$$\delta|y, \beta, \tau, \Lambda, z_1, \ldots, z_n \sim$$
$$N\left\{\left(\tau\sum_{i=1}^{n}\lambda_i z_i^2 + \frac{1}{\sigma_\delta^2}\right)^{-1}\left(\tau\sum_{i=1}^{n}\lambda_i z_i\hat{y}_i + \frac{\mu_\delta}{\sigma_\delta^2}\right), \left(\tau\sum_{i=1}^{n}\lambda_i z_i^2 + \frac{1}{\sigma_\delta^2}\right)^{-1}\right\}, \tag{9}$$

where μ_δ and σ_δ^2 are the conjugate Normal prior parameters for δ.

For the λ_i, it can be shown that

$$\lambda_i|y, \beta, \tau, \nu, \delta, z_1, \ldots, z_n \sim \Gamma\left\{\frac{1}{2}(\nu + 1), \left[\frac{1}{2}\left(\tau\tilde{\tilde{y}}_i^2 + \nu\right)\right]^{-1}\right\}, \tag{10}$$

with the skewness built into the distribution by replacing $\hat{y}_i$ with $\tilde{\tilde{y}}_i$.

The posterior for ν, conditional on Λ, and its priors, are given in the following equations.

$$p(\nu|y, \Lambda) \propto \frac{\nu^{\frac{1}{2}\nu n}}{2^{\frac{1}{2}\nu n}\left[\Gamma\left(\frac{\nu}{2}\right)\right]^n}|\Lambda|^{\frac{1}{2}\nu-1}\exp\left[-\frac{1}{2}\nu\sum_{i=1}^{n}\lambda_i\right]p(\nu), \tag{11}$$

with the prior on ν taking one of four forms, namely the truncated exponential, the Independence Jeffrey's prior, the probability matching prior or reference priors for the orders (ν, μ, σ^2), (ν, σ^2, μ),

and $\left(\mu, \nu, \sigma^2\right)$, and the reference priors for the orders $\left(\mu, \sigma^2, \nu\right)$, $\left(\sigma^2, \mu, \nu\right)$, and $\left(\sigma^2, \nu, \mu\right)$. In this paper the Independence Jeffrey's prior is used. We find that this prior is less restrictive on the degrees on freedom than the well-established exponential prior, within the context of the strobit estimation. It is shown by Fonseca et al. (2008) that the independence Jeffreys prior is

$$p_{IJEFF}\left(\nu, \beta, \sigma\right) \propto \sigma^{-1}\left(\frac{\nu}{\nu+3}\right)^{\frac{1}{2}}\left[\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2\left(\nu+3\right)}{\nu\left(\nu+1\right)^2}\right]^{\frac{1}{2}}$$

assuming that the marginal priors for β and (σ, ν) are independent *a priori*, and $\psi'(\cdot)$ is the trigamma function.

Working with the natural log posterior and log priors is easier:

$$\begin{aligned}\log\left(p\left(\nu|y,\lambda\right)\right) \propto & \frac{1}{2}\nu n\log\left(\nu\right) - \frac{1}{2}\nu n\log\left(2\right) - n\log\left(\Gamma\left(\frac{\nu}{2}\right)\right) \\ & - \left(\frac{1}{2}\nu - 1\right)\sum_{i=1}^{n}\log\left(\lambda_i\right) - \left(\frac{1}{2}\nu - 1\right)\sum_{i=1}^{n}\lambda_i - \log\left[p_{IJEFF}\left(\nu,\beta,\sigma\right)\right]. \qquad (12)\end{aligned}$$

The algorithm for the Gibbs sampler (and Metropolis sampler for ν) when we wish to incorporate skewness into the imputation model utilises the conditional distributions listed above.

3.2. Strobit and ordered strobit model estimation

Now that we have the conditional distributions of the parameters of the skew t-distribution, we can use these draws in the place of the draws of probit parameters, namely the β.

Once more, after initialising the parameters, if the strobit is estimating a two-category response variable, the category split on the latent variable is set at 0. Otherwise, similarly to the ordered probit estimation, the first category split, γ_1, is set at 0, while the remaining category splits, the γ_j, become parameters to be estimated in the same way as for the ordered probit, namely their conditional distribution follows Equation (3).

Given all the other unknowns, we can draw bounded latent variables as follows:

$$\begin{aligned} W_i|y,\beta,\tau=1,\delta,Z,\nu &\sim t_\nu + X\beta + Z\delta \\ &\quad \text{truncated at the left by 0 if } y_i = 2 \\ W_i|y,\beta,\tau=1,\delta,Z,\nu &\sim t_\nu + X\beta + Z\delta \\ &\quad \text{truncated at the right by 0 if } y_i = 1 \qquad (13)\end{aligned}$$

Thus, for a Gibbs sampler to simulate draws from the joint posterior we us the following algorithm:

1. Initialise β using the Gibbs sampler with y as the dependent variable, follow up with initialisation of all the other parameters.
2. Set the category split at a latent value of 0 in the case of 2-category response variable, or, in the case of the ordered response variable, draw the γ_j variables from Equation (3), given the preceding draw of W.
3. Generate a vector W from Equation (13).
4. Step through the Gibbs sampler for conditional draws from Equations (6)-(11), each paramater based on preceding draws of the other parameters.

5. Repeat steps 2–4 until the draws for W and the parameters converge.

3.2.1. Some practicalities

If we follow the algorithm above, then the draws for the parameters (β in particular) vary widely from round to round. Theoretically, the draws should be stable, but the variance in the draws makes any prediction based on a single draw in step 4 rather unreliable. In order to stabilise the draws, step 4 of the above procedure is repeated several times, say 200 times, until a conservative set of draws for the parameters of the skew t-distribution is evident.

While this modification of the algorithm considerably increases its running time, the modification is necessary if the fitting algorithm is to be used for prediction. In prediction, only a single draw from the end of the Gibbs sampler is used, and if the variation from one draw to the next is very high, one is likely to obtain drastically different coefficient estimates from one run of the fitting procedure to the next.

Through thorough investigation, we are satisfied that this extra smoothing step does not detract from the implementation of the strobit model except in the case where there are time constraints for the fitting procedure.

4. Simulation Methodology

4.1. Simulated data

In order to assess the robustness of the probit and strobit models, and their ordered counterparts, four difference latent data construction scenarios are examined: Normal, skew t, Exponential and Uniform. We assume $U = -1 + 4x + \xi$, where U is the true latent variable, $x \sim N(0.5, 1)$, and ξ is an error that depends on the data scenario under question:

1. Normal data: $\xi_i \sim N(0, 1), i = 1, \ldots, n$;
2. Skew t data: $\xi_i = -2z_i + 0.5w, i = 1, \ldots, n$, where $z_i \sim N(0, 1)I_{z_i>0}$, in other words, the z_i are positively truncated Normal random variables, and $w_i \sim t_5$;
3. Exponential data: $\xi_i \sim Exp(1), i = 1, \ldots, n$ or $\xi_i = -\ln(1 - u_i)$, where $u_i \sim UNF(0, 1)$;
4. Uniform data: $\xi_i, i = 1, \ldots, n$, is a random integer between 0 and 5.

Once the latent data is generated, the observations are allocated to categories based on this observed latent data using random category splits in the full simulation analysis, or splitting point(s) -2 (and 2) for the 2-category (3 category) single simulation discussion.

Two sample sizes are considered in the full simulation study, namely $n = 200$ and $n = 1000$, but for the single simulation analysis, the review is restricted to $n = 1000$.

4.2. Assessment Methods

The primary method of assessing the probit and strobit models, as well as their ordered counterparts, is using the mean absolute deviation (MAD) of the predicted category values from their

actual category values given a new sample for a particular data scenario. This is essentially a summary of the classification matrices across multiple simulations within each data scenario. In brief, we proceed using the following steps:

1. Generate latent data dependent on an exogenous Normal random variable, x, an intercept, and an error appropriate to the data scenario under examination.
2. Split the latent data at random points to generate a categorical variable (ensuring that each category contains at least 2% of the sample).
3. Estimate parameter values for the (ordered) probit and strobit models on the given simulated data, using the average of 300 draws from the Gibbs sampler, after a burn-in of 300 draws. Within the strobit estimation, the smoothing process also burns in 50 draws of the skew t-distribution parameters within each of the 600 strobit Gibbs sampler runs.
4. Generate a new sample according to the same latent data scenario of step 1.
5. Using the random splits generated in step 2, re-split the new sample into categories.[2] These categories are the 'correct' categories for the new sample.
6. Using the estimated parameter values for the regression model estimated in step 3, predict a latent value for each observation in the new sample, drawing random Normal errors for the probit and ordered probit predictions, and skew t errors for the stobit and ostrobit models. [3]
7. Using the latent predictions and the estimated category splits from the model estimation, re-categorise the new sample. These categories are the predicted categories of the new sample.
8. Calculate the MAD for a model by averaging the absolute difference between actual and predicted categories of the new sample.
9. Repeat steps 1–8 for a total of 200 simulations.

5. Simulation Analysis

In this section, a single simulation across all data scenarios is scrutinised, and then the process is repeated for a total of 200 runs for a thorough assessment of the methodology.

5.1. Single-run analysis

In order to understand the simulation analysis, the histograms of the data, as well as histograms for the errors that are added to the exogenous covariates, are presented in Figures 1 and 2, for a two- and three-category simulation, respectively. From these figures, it is clear that the latent data is modelled as a regression on a Normal covariate and an intercept, and is coupled with varying errors, including Normality (scenario 1), negative skewness (scenario 2), positive skewness (scenario 3), and uniformity (scenario 4). The probit and strobit Gibbs sampler draws for the two-category model estimation (after burn-in) are shown in Figures 3 and 4. These parameter draws are particularly

[2] It can be noted that in some instances, this procedure led to one category containing all the observations. These cases were not eliminated, since the model could still theoretically predict an observation outside of the category bounds containing all these observations, leading to classification error.

[3] Symmetric t errors combined with a aero-truncated Normal error for skewness

stable, except for the degrees of freedom, ν, for the strobit estimation. The probit and strobit Gibbs sampler draws for the three-category model estimation (after burn-in) are shown in Figures 5 and 6. One will notice in these figures that there is sometimes drift in both the γ_1 value and a β value. This drift is not much of a concern as long as the draws drift together — one cannot pin more than one category boundary down without severely limiting the estimation procedure. One might argue that for three categories, one could fix the category boundaries and hope that the sampler is long enough to squeeze and move the underlying latent model to correctly fit the data, but beyond three categories this would be unrealistically strict. In any case, the drift of the strobit parameter pairs is not entirely a problem, since we are not using the estimation procedure for interpretation of fit parameters, but merely for prediction (and classification). Parameter pair drift will not affect this goal.

Once the probit and strobit models are fitted under each data scenario, the fitted latent distributions are graphed in Figure 7 for two categories and Figure 8 for three categories. The different shades indicate the different sequentially observed categories. Note that the fitted latent data is forced to be separated by category, leading to multi-modal distributions. One would hope that the estimation algorithms would lead to smooth, uni-modal fitted distributions, but this is not the case, even for the probit on Normal data.

Once the models are estimated, a new sample is drawn according to the appropriate data scenario, and the estimated models are used to predict a new distribution of the latent data. Histograms of these distributions are given in Figure 9 for two categories, and Figure 10 for three categories, and are shaded according to the categories that the new sample's observations would be assigned to had the underlying model been known. It is clear from these figures that there is no way of splitting all the new observations using their predicted latent data into their correct categories. This leads to classification error. A visual representation of the classification matrix for the three-category simulation is given in Figure 11.

For the two simulations represented in the graphs, we have the following classification errors for the new samples: for two categories, the probit has MAD errors of 18%, 18.8%, 20.6% and 14.9% for the Normal, skew t, Exponential and Uniform data scenarios respectively, while the srobit has MAD errors of 13%, 9.9%, 13.7% and 15.6% for the four data scenarios, respectively; for three categories the probit has MAD errors of 32.1%, 37.7%, 27.2% and 41% for the Normal, skew t, Exponential and Uniform data scenarios respectively, while the srobit has MAD errors of 21.6%, 24.5%, 19.6% and 28% for the four data scenarios, respectively. These figures have little value without repeating the simulation process multiple times, as is carried out in the next section.

5.2. Multiple-run analysis

The initial simulation analysis, summarised in Table 1, seems promising for the strobit model. In all simulation scenarios, across two and three categories, sample sizes of both 200 and 1000, and across all four data scenarios, the strobit model's MAD error is more often than not lower than that of the probit model's MAD error.

However, upon further analysis, the strobit model loses some of its favour. The first problem becoming evident is the number of times within the multiple simulation procedure that MAD errors from the probit and strobit models are the same. In Figures 12 – 15, we plot the difference between probit and strobit MAD errors against a measure of tail-category sparseness or observation-scarcity,

Table 1: MAD error superiority proportions, by category, sample size, and data scenario

Categories	Sample size	Data scenario	Probit better	Models equal	Strobit better
2	**200**	Normal	20.5%	33.5%	46.0%
		skew t	12.5%	41.0%	46.5%
		Exponential	16.5%	40.5%	43.0%
		Uniform	15.5%	43.0%	41.5%
	1000	Normal	28.0%	27.5%	44.5%
		skew t	18.5%	31.0%	50.5%
		Exponential	18.5%	38.0%	43.5%
		Uniform	20.5%	39.0%	40.5%
3	**200**	Normal	20.5%	33.5%	46.0%
		skew t	12.5%	41.0%	46.5%
		Exponential	16.5%	40.5%	43.0%
		Uniform	15.5%	43.0%	41.5%
	1000	Normal	41.0%	4.5%	54.5%
		skew t	39.5%	7.0%	53.5%
		Exponential	42.5%	9.5%	48.0%
		Uniform	29.5%	8.5%	62.0%

namely the negative sum of the natural logs of the proportions of observations in the tail categories.[4] We find that in the two-category case the probit and strobit models are misclassifying the same proportions when tail scarcity is high, *i.e.* the strobit model is not doing better than the probit in classifying observations into the correct categories when those categories are sparsely-populated tail categories. This is quite a concern, since the strobit model, with an underlying heavy-tailed skew distribution, might naturally be thought of as more capable in this context.

Another issue becomes apparent when one examines the difference in MAD errors between the probit and strobit models across multiple simulations: this difference is not significantly greater than zero, *i.e.* while it is evident that the strobit model's MAD error is often smaller than that of the probit, the average strobit MAD error is not significantly lower than that of the probit. This is illustrated in Figures 16 – 19. The empirical 95% intervals for the probit minus strobit MAD error crosses over zero for all data scenarios under both the two- and three-category cases.

Apart from a general review of the analyses performed, one can note a few interesting results from this study. It is clear that the strobit model performs better classifications that the probit when the latent data is Normally distributed. This is a strange result, but can be partially explained by the fact that the Normal distribution is a special case of the skew t. Also, for the two-category scenarios, the strobit does perform better than the probit model when the scarcity measure is not extremely high, but moderately large (Figures 12 and 14).

6. Conclusion

This paper introduces a new robust Bayesian procedure for modelling ordinal categorical response data as a function of exogenous covariates. The work is based on the Bayesian estimation processes of probit, as explained by Albert & Chib (1993), the logit, as explained by Groenewald & Mokgatlhe (2005), and the robit, as explained by Liu (2005). The modelling procedure expands on the existing

[4] Or simply negative sum of the natural logs of the proportions of observations in both categories in the two-category case.

literature by assuming that the (ordinal) categorical responses are linked to skew t-distributed latent data — this model is then called the strobit model. Procedures are introduced to estimate the parameters of this Bayesian strobit model. Since the strobit model fits more parameters than the probit and logit, the estimation procedure can be quite time consuming, and some practicalities associated with this process are discussed. It is noted that the Bayesian estimation procedures of categorical response models such as the probit, logit and strobit, produce parameters that are linked to unknown underlying latent data, and are thus not useful for interpretation, but only for prediction or classification of new observations.

The probit and strobit model are compared under two-category and three-category binary responses based on simulated latent data with various characteristics. The strobit model performs marginally better than the probit under all data situations (even when the latent data is Normally distributed), but the difference in performance between the two models is not significant. However, since the Normal distribution is a special case of the skew t-distribution, and hence the probit is a special case of the strobit, and since the strobit performs marginally better than the probit model under varying data scenarios (including Normality), the authors recommend that if computing time is not of great concern to a modeller, the Bayeian estimation of the strobit model be used in place of the Bayesian estimation of the probit.

Naturally, this study opens up various topics for further research. The strobit model is built for implementation in sequential regression multiple imputation (SRMI), and thus further research into the applicability of this model in that context is warranted. Moreover, this model should be compared with other categorical imputation procedures, such as the multinomial model that is commonly used in SRMI, or other categorical response models. Also, the fact that the strobit model, as it is defined in this paper, does not significantly improve the classification results on tail categories with low counts, is a concern. Perhaps a skew t model with more allowance for skewness could be examined (*i.e.* $z_i \sim t_3 I_{z_i>0}$ instead of $z_i \sim N(0,1) I_{z_i>0}$ could be used). As far as the estimation procedure itself is concerned, work should be done to speed up the Gibbs sampler, as well as stabilise the actual sampling. This research required sampling-in-sampling to obtain stabilised parameter estimates, and research can be done on the suitability and efficiency of such a procedure.

References

Albert, J. & Chib, S. (1993), 'Bayesian analysis of binary and polychotomous response data', *Journal of the American Statistical Association* **88**(422), 669–679.

Fonseca, T. C. O., Ferreira, M. A. R. & Migon, H. S. (2008), 'Objective bayesian analysis for the student-t regression model', *Biometrika* **95**(2), 325–333.

Groenewald, P. & Mokgatlhe, L. (2005), 'Bayesian computation for logistic regression', *Computational Statistics and Data Analysis* **48**, 857–868.

Liu, C. (2005), Robit regression: A simple robust alternative to logistic and probit regression, *in* A. Gelman & X.-L. Meng, eds, 'Applied Bayesian Modeling and Causal Inference from Income-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family', John Wiley & Sons, Chichester, U.K., pp. 227–238.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. & Solenberger, P. (2001), 'A multivariate technique for multiply imputing missing values using a sequence of regression models', *Survey Methodology* **27**(1), 85–95.

Appendix A Graphs

Figure 1: Latent data under the four data scenarios for the two-category analysis, with embedded errors; $n = 1000$.

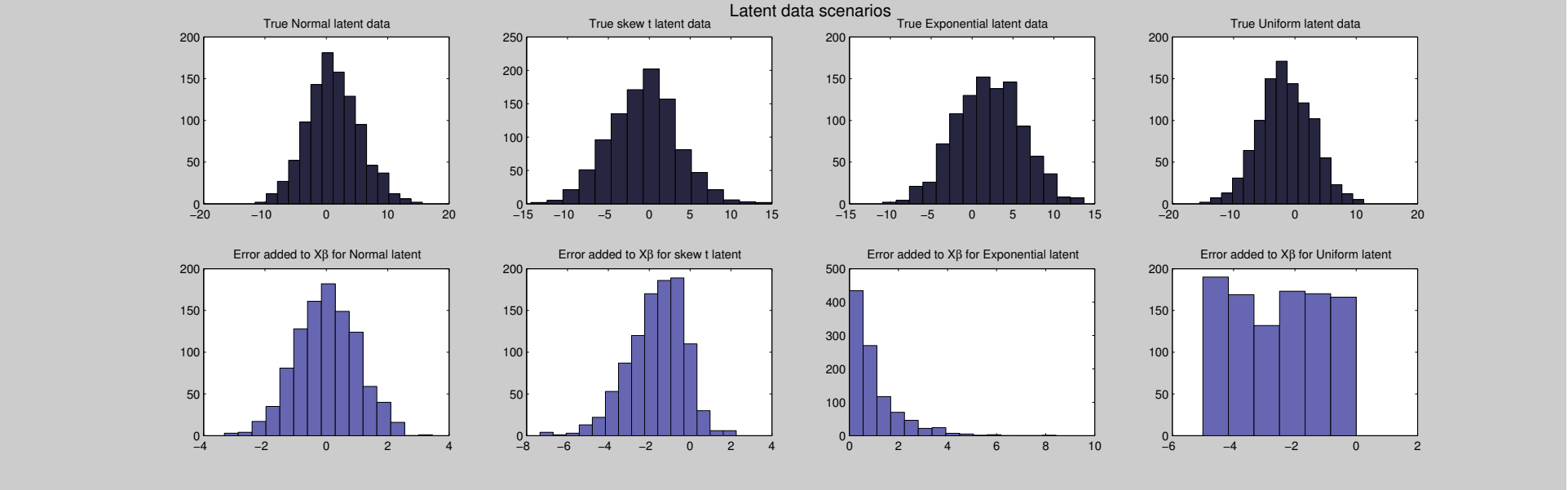


Figure 2: Latent data under the four data scenarios for the three-category analysis, with embedded errors; $n = 1000$.

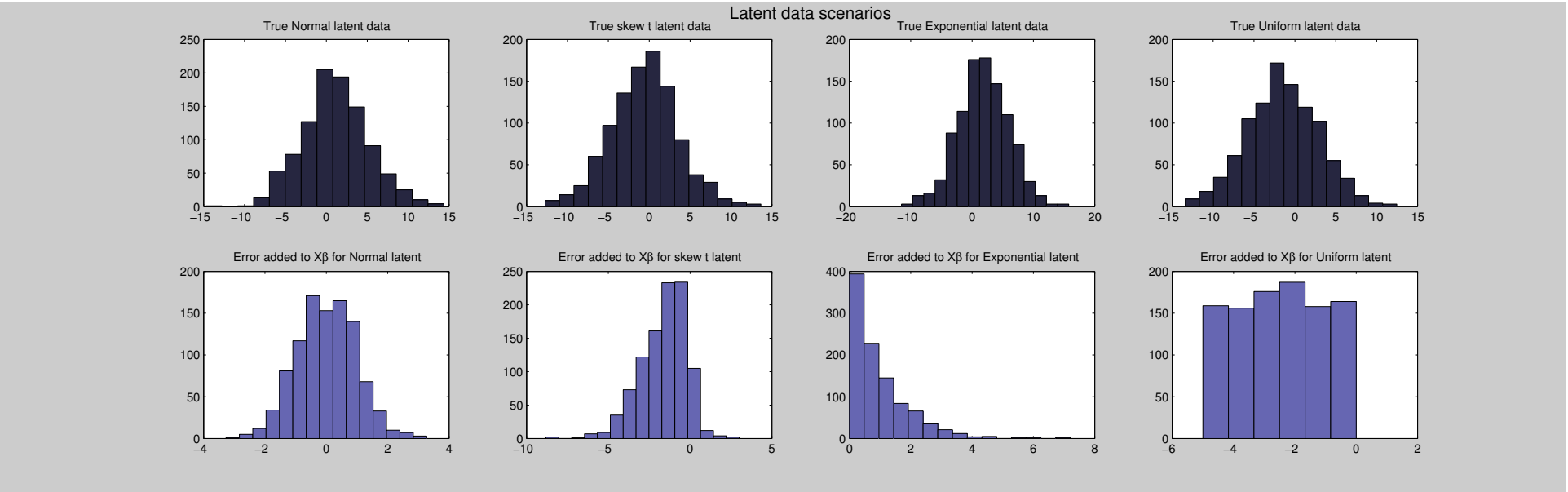


Figure 3: Probit Gibbs sampler draws after burn-in for the two-category data

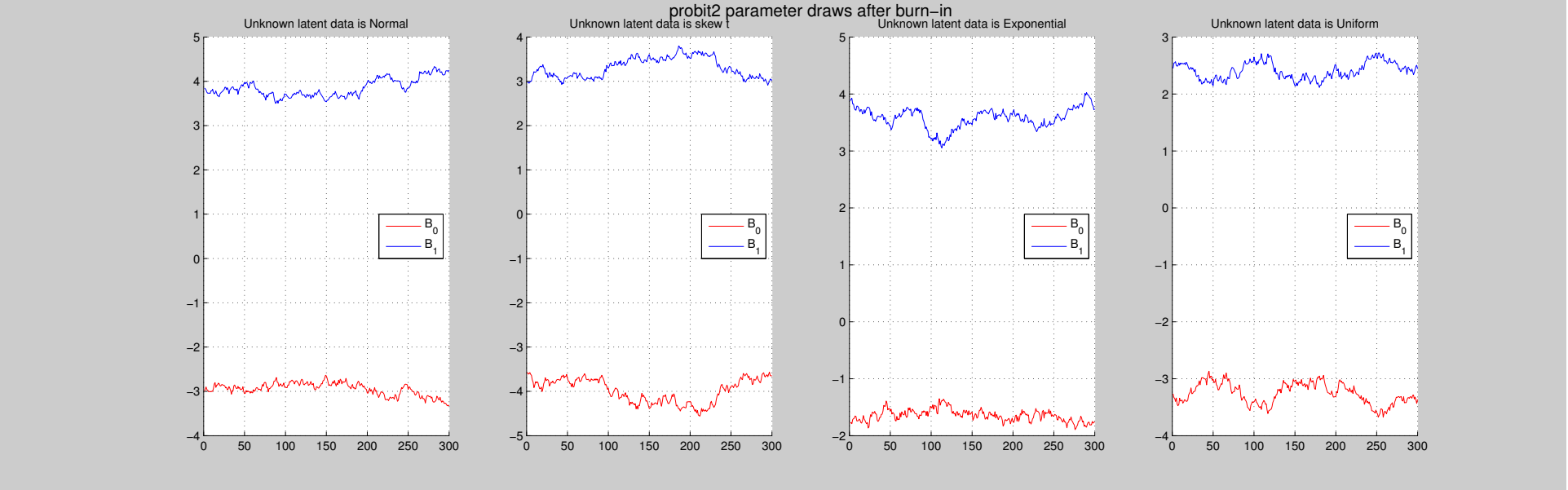


Figure 4: Strobit Gibbs sampler draws after burn-in for the two-category data

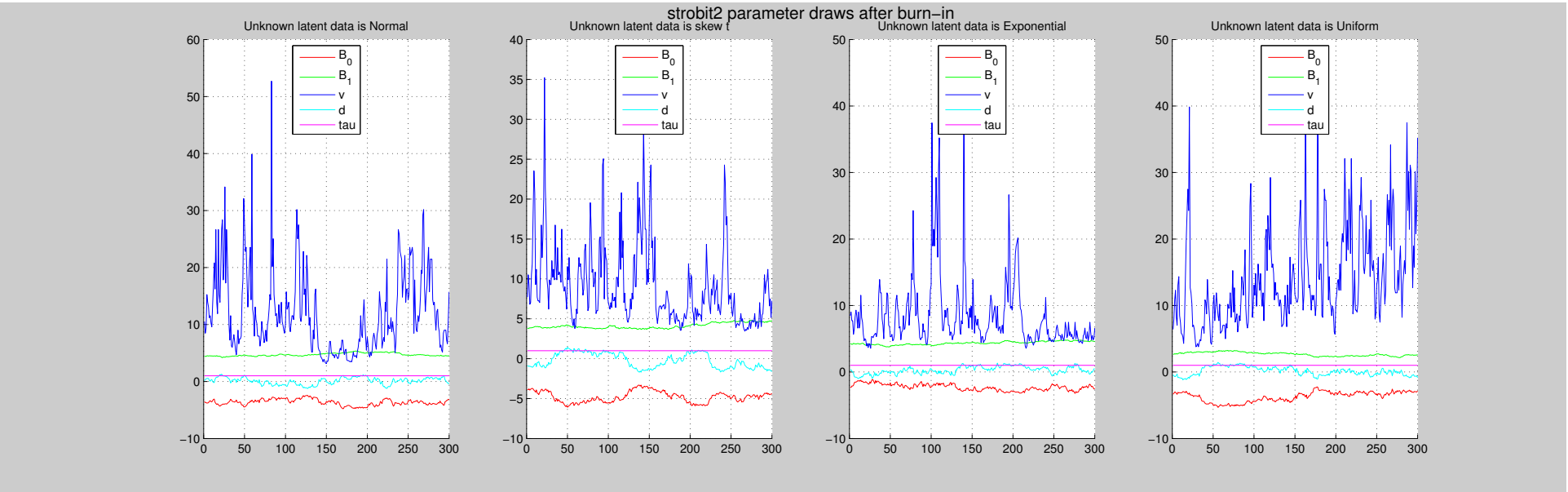


Figure 5: Probit Gibbs sampler draws after burn-in for the three-category data

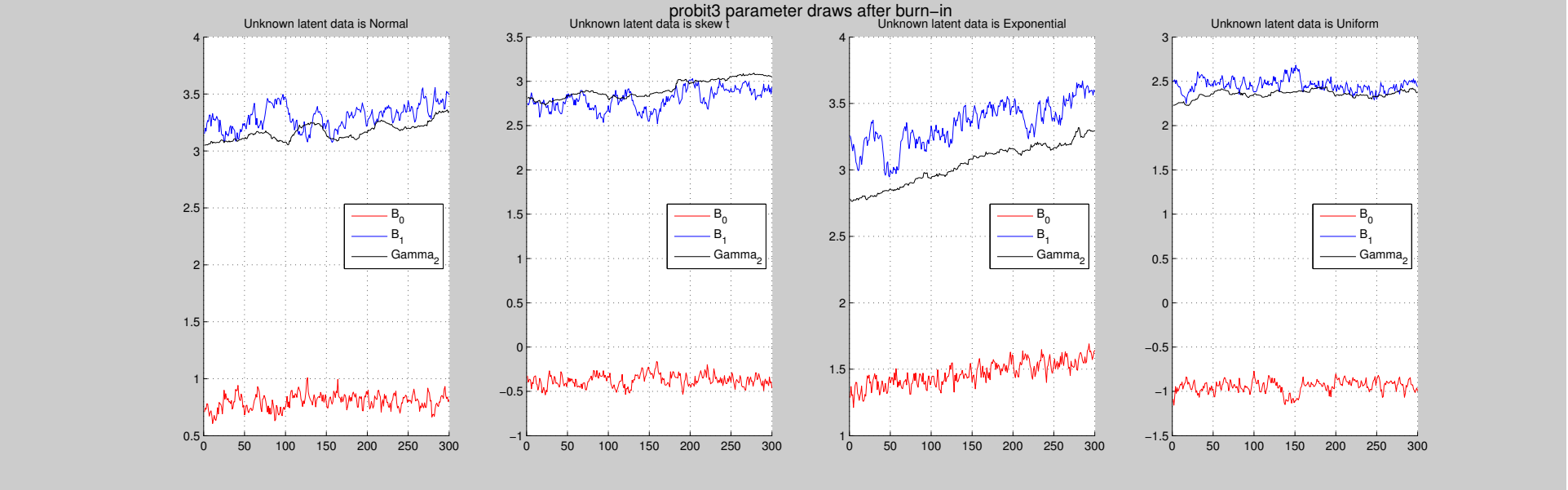


Figure 6: Strobit Gibbs sampler draws after burn-in for the three-category data

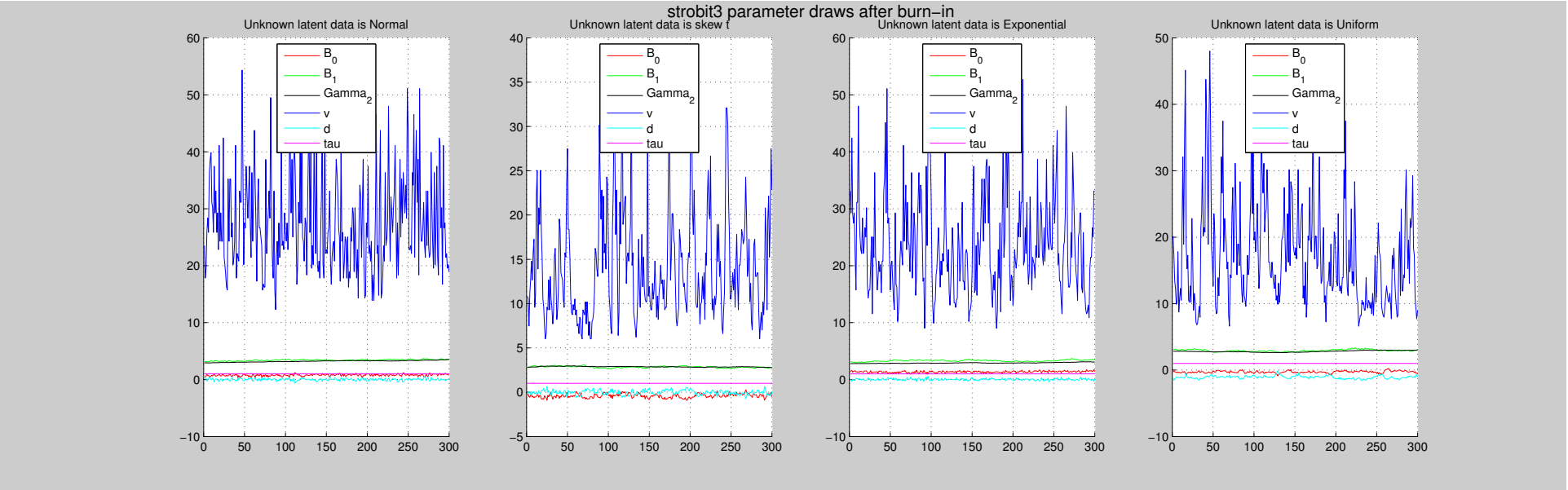
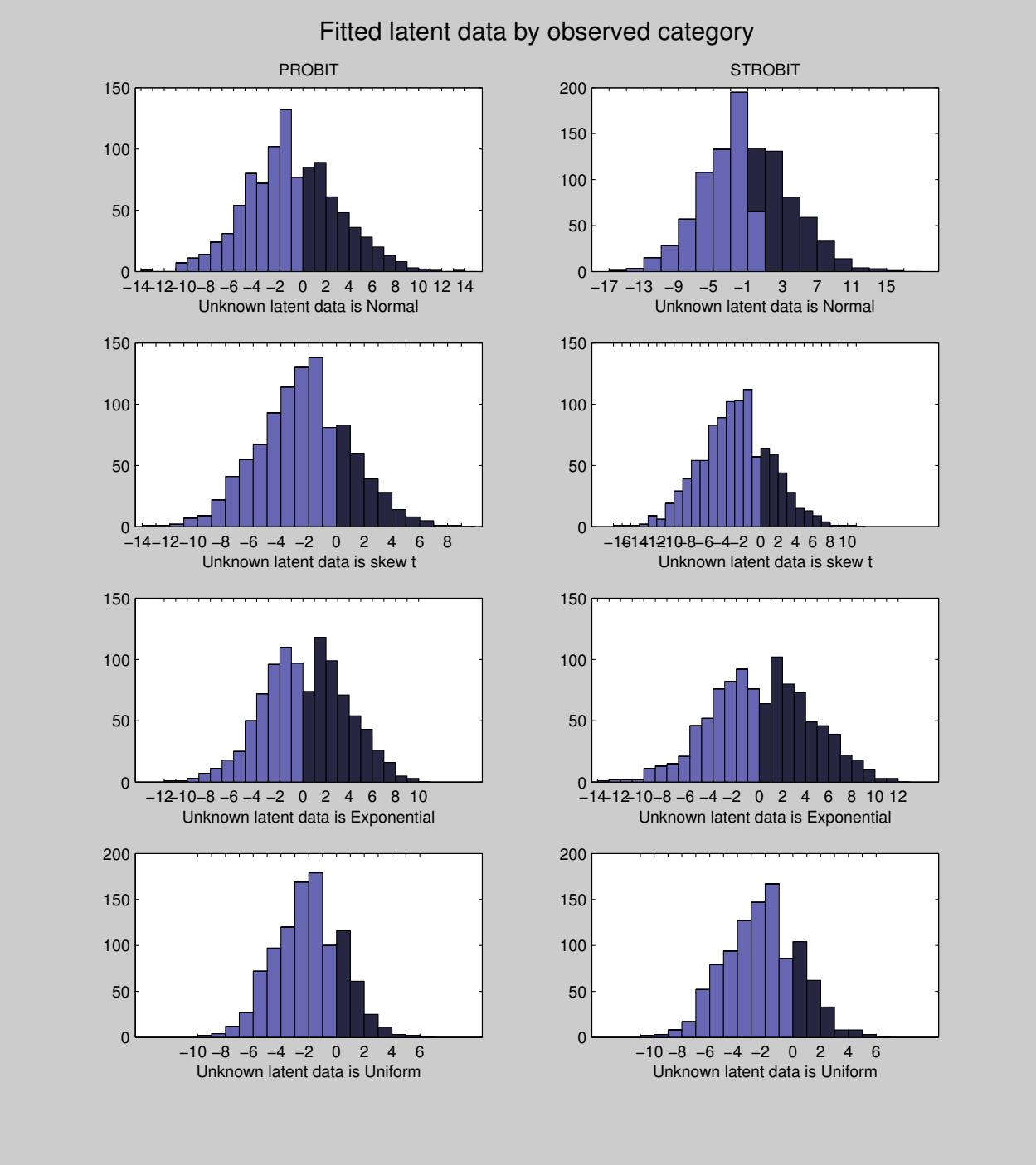
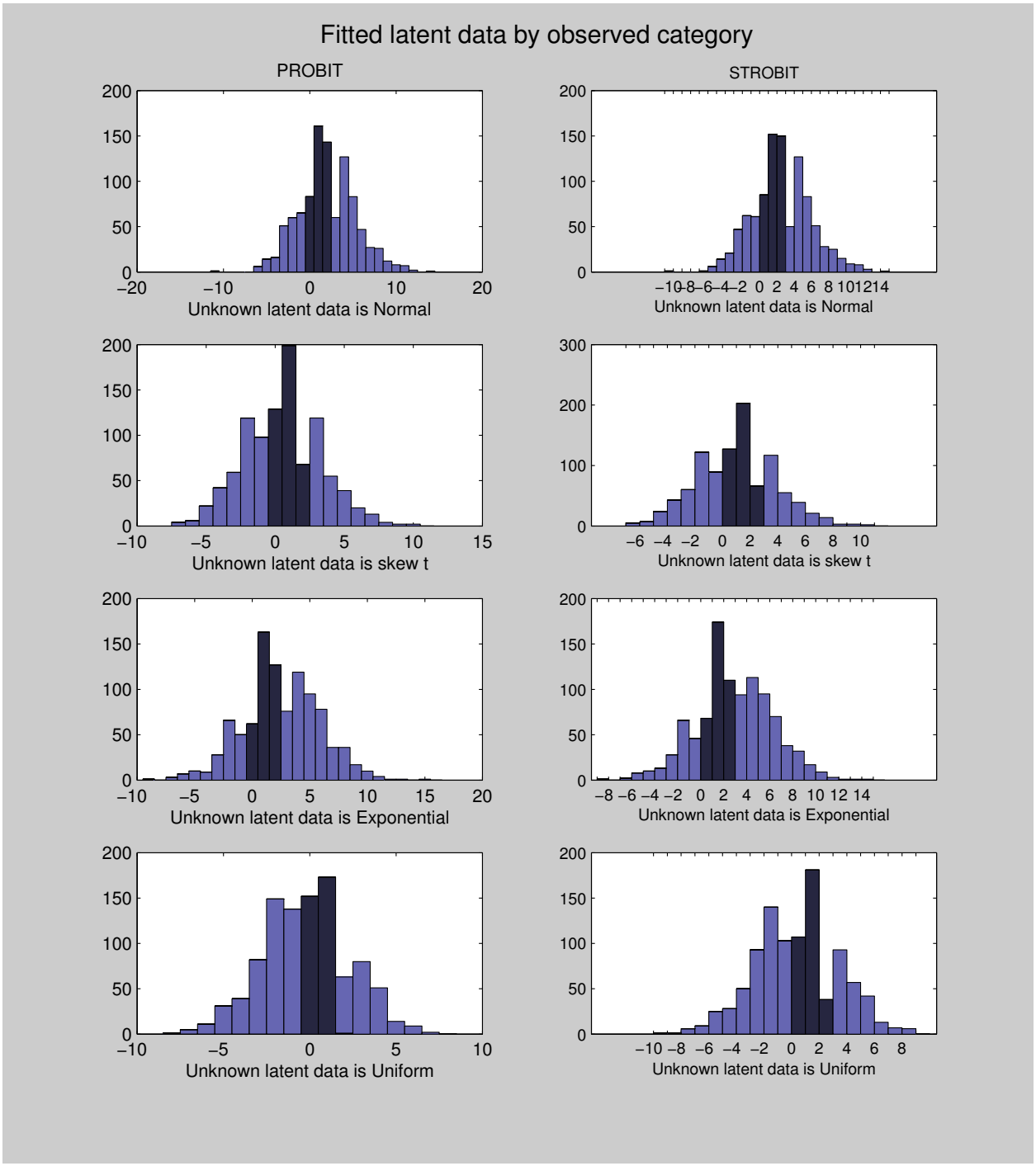


Figure 7: Fitted latent data (2 categories)



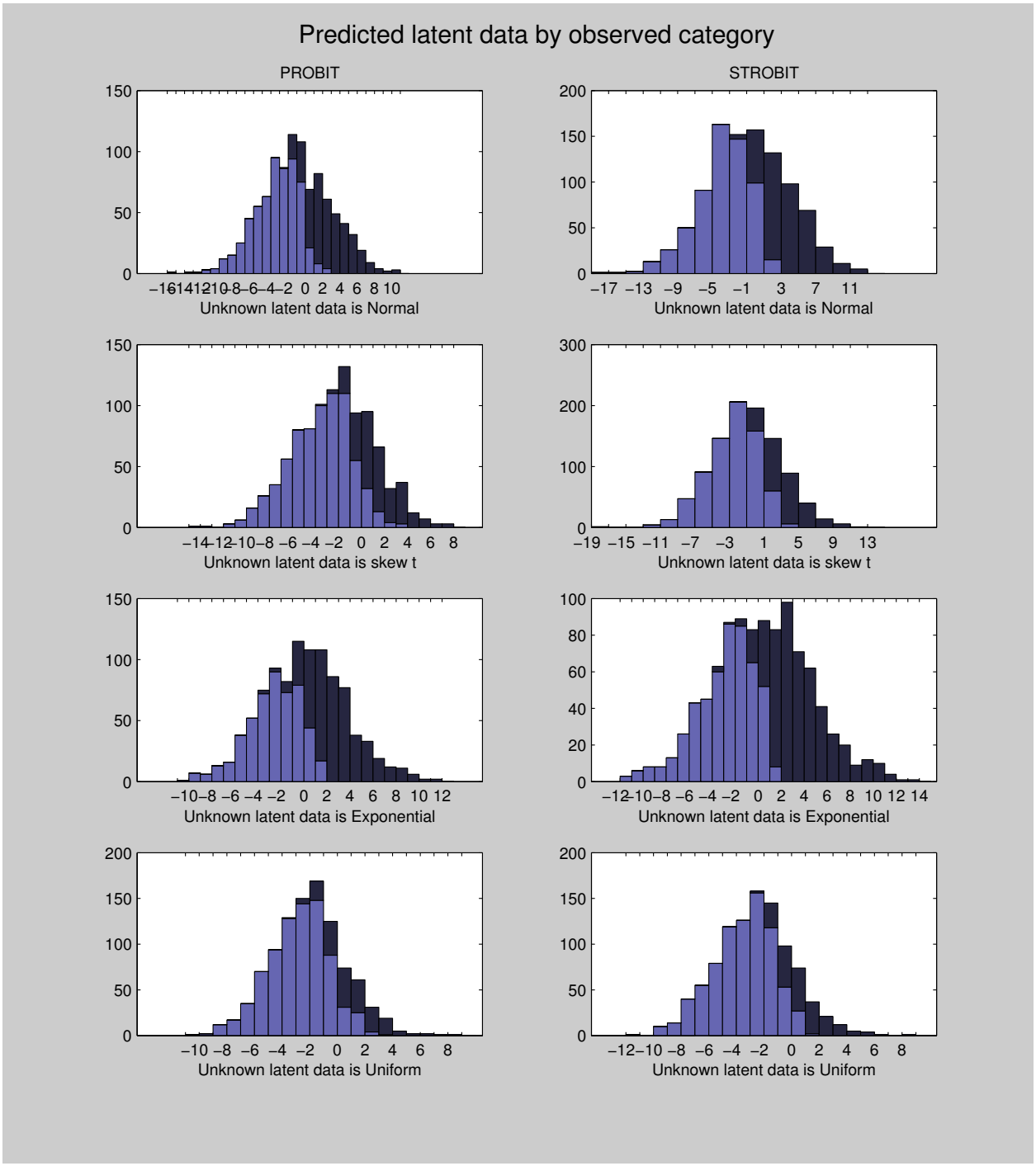
The different shades represent the two different observed categories. The latent data is separated for each category by the fitting algorithm.

Figure 8: Fitted latent data (3 categories)



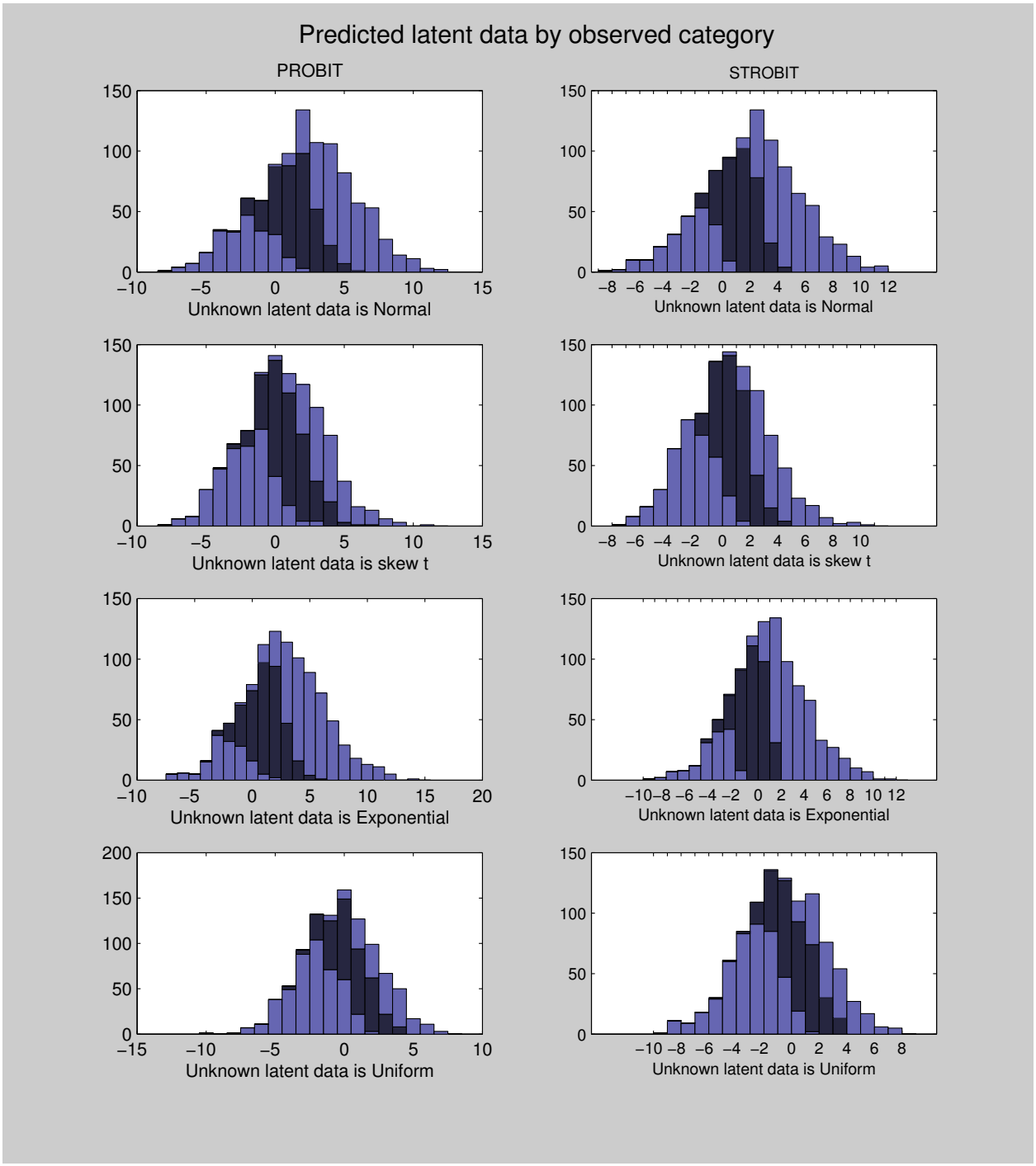
The different shades represent the three different observed categories. The latent data is separated for each category by the fitting algorithm.

Figure 9: Predicted latent data (2 categories)



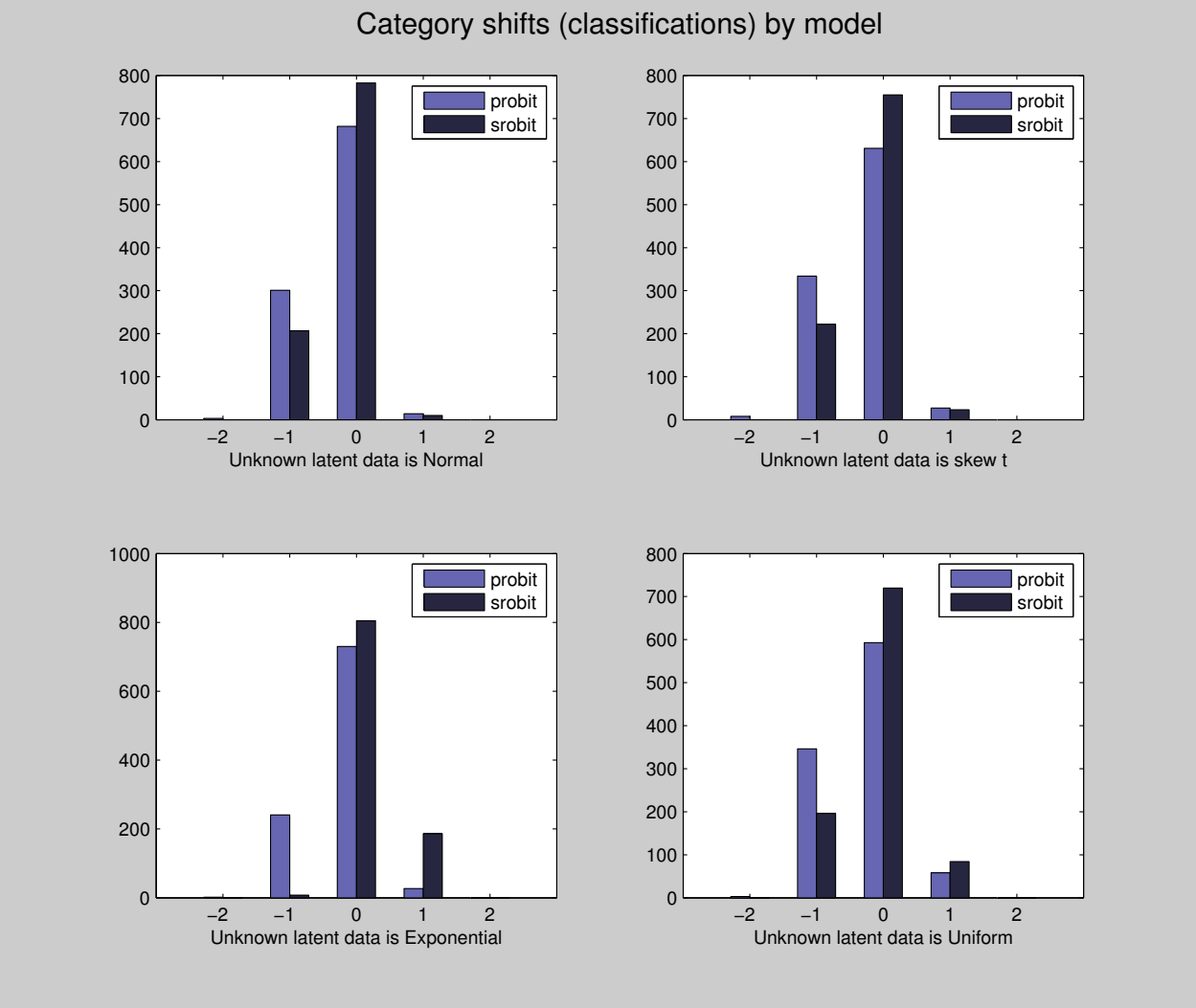
The different shades represent the two different actual observed categories, and not the categories that are chosen off the predicted latent data.

Figure 10: Predicted latent data (3 categories)



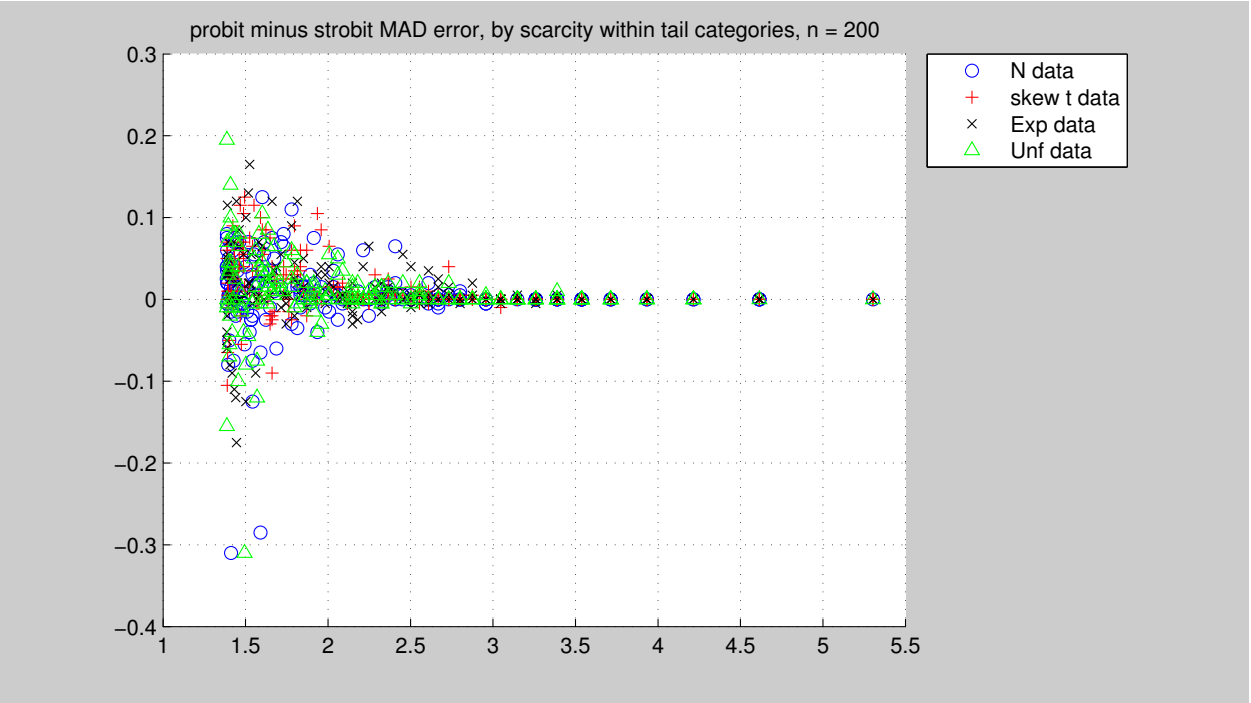
The different shades represent the two different actual observed categories, and not the categories that are chosen off the predicted latent data.

Figure 11: Classification errors for three-category simulation



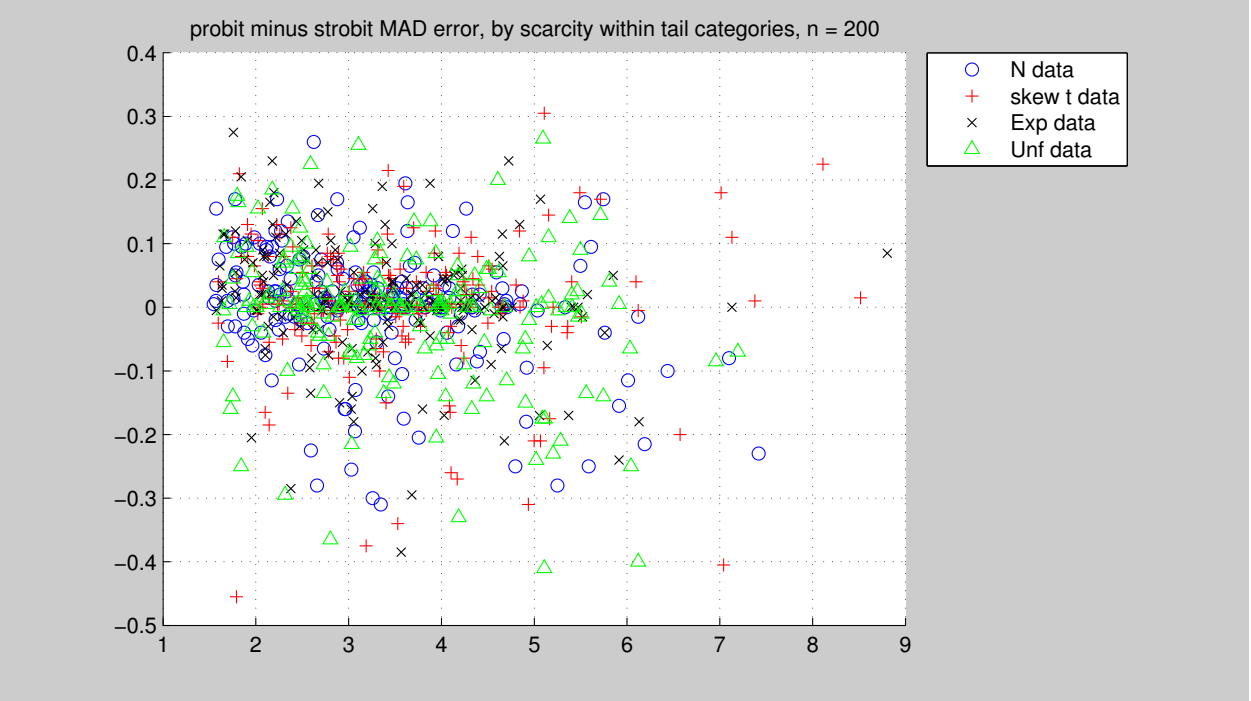
This set of histograms represents a visual take on a classification matrix. Off-diagonal elements of the classification matrix (*i.e.* classification errors) are to the right and left of the '0' column.

Figure 12: Two-category MAD error difference on category sparseness, by data scenario, $n = 200$



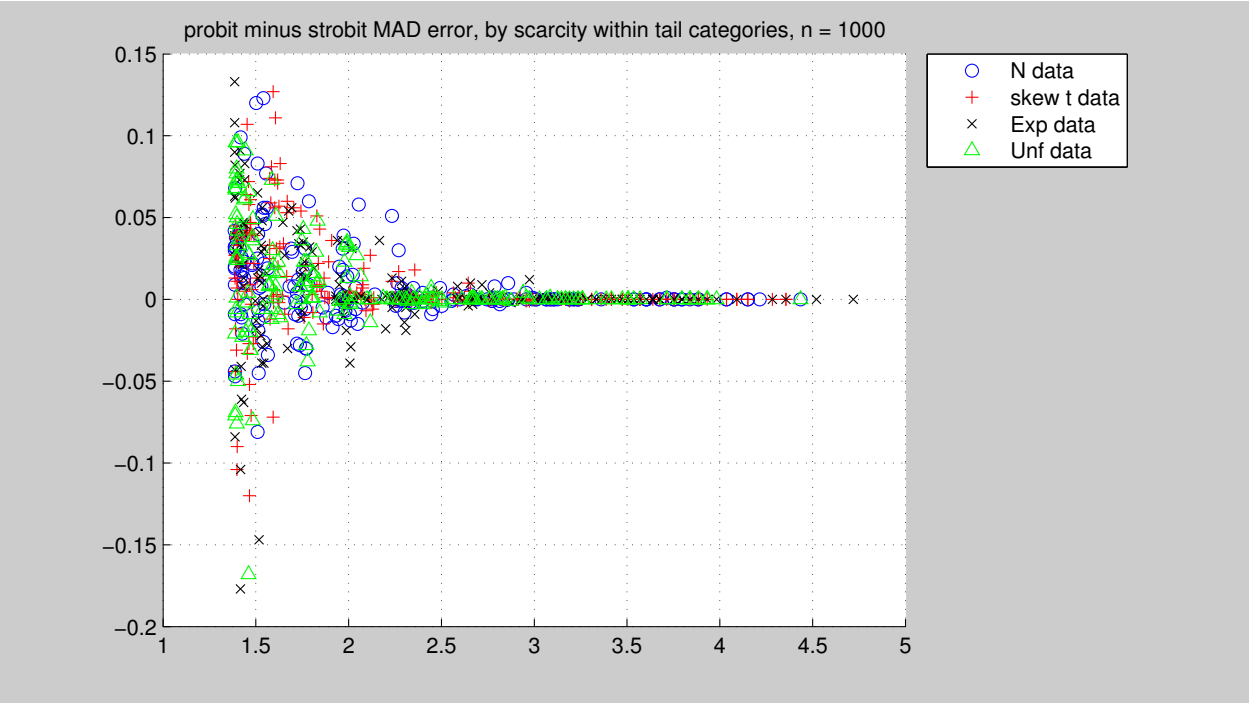
Scatterplot of probit minus strobit MAD error by negative sum of logs of category proportions for the two-category multiple simulation procedure.

Figure 13: Three-category MAD error difference on tail category sparseness, by data scenario, $n = 200$



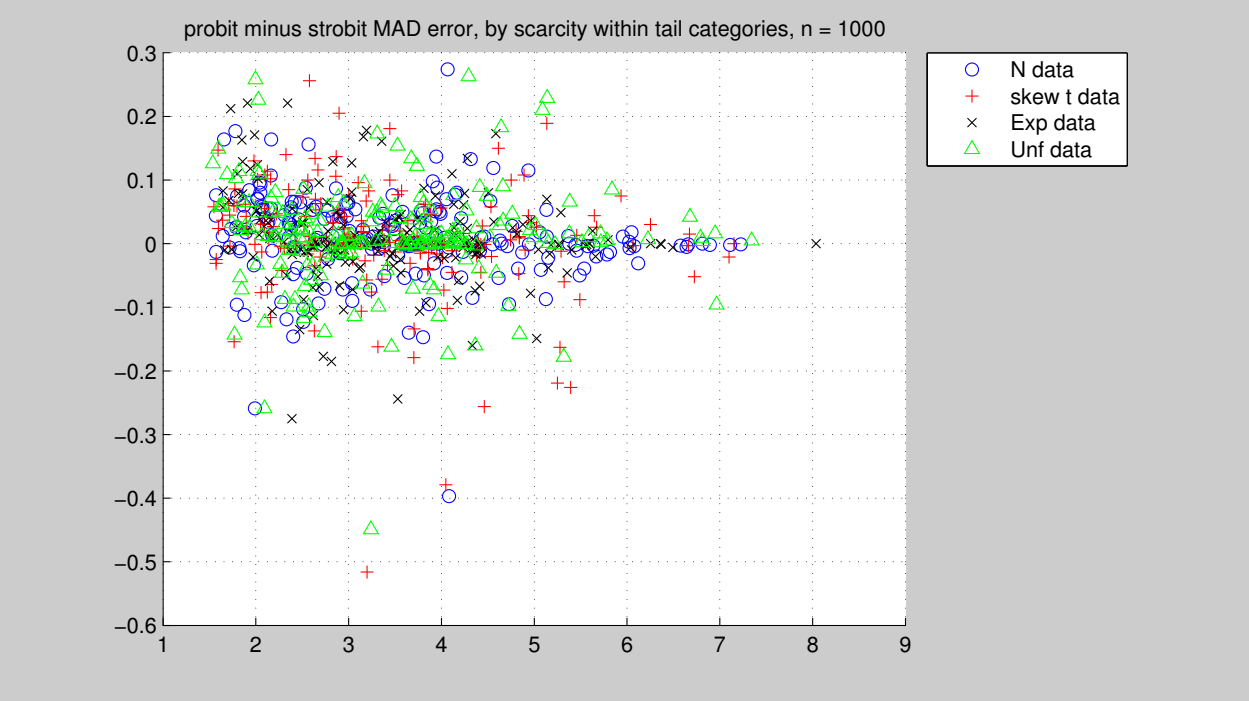
Scatterplot of probit minus strobit MAD error by negative sum of logs of outer category proportions for the three-category multiple simulation procedure.

Figure 14: Two-category MAD error difference on category sparseness, by data scenario, $n = 1000$



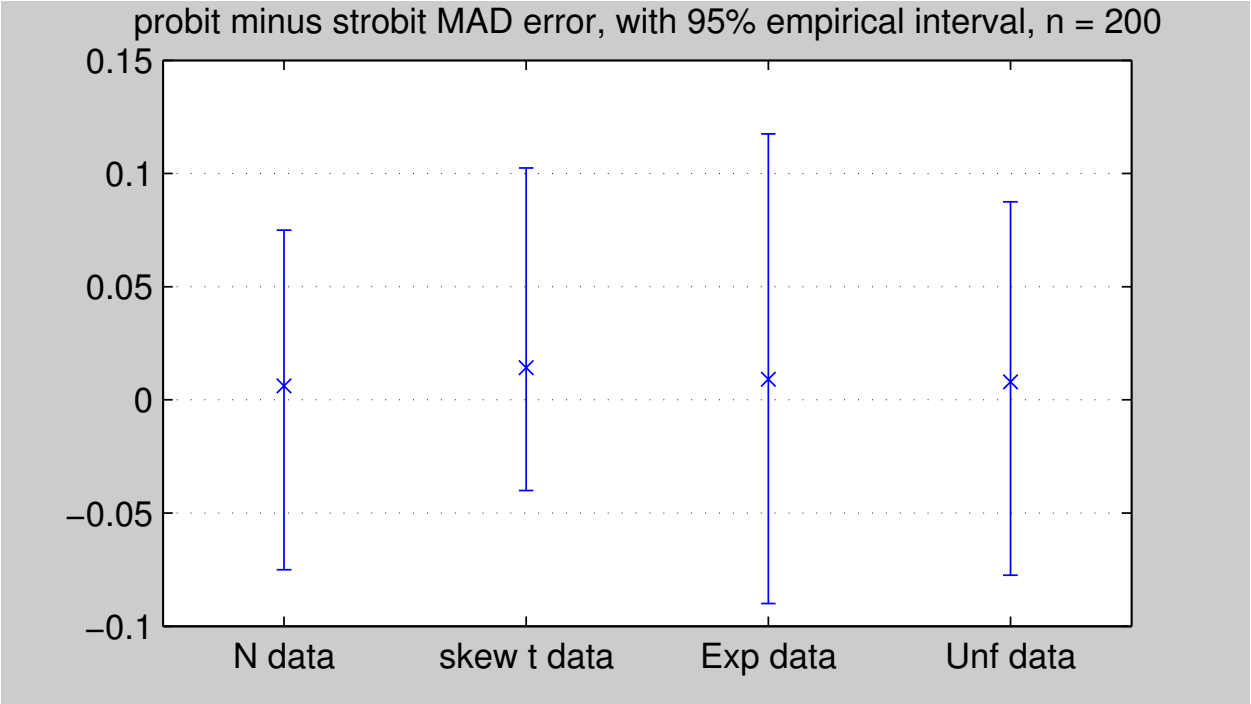
Scatterplot of probit minus strobit MAD error by negative sum of logs of category proportions for the two-category multiple simulation procedure.

Figure 15: Three-category MAD error difference on tail category sparseness, by data scenario, $n = 1000$



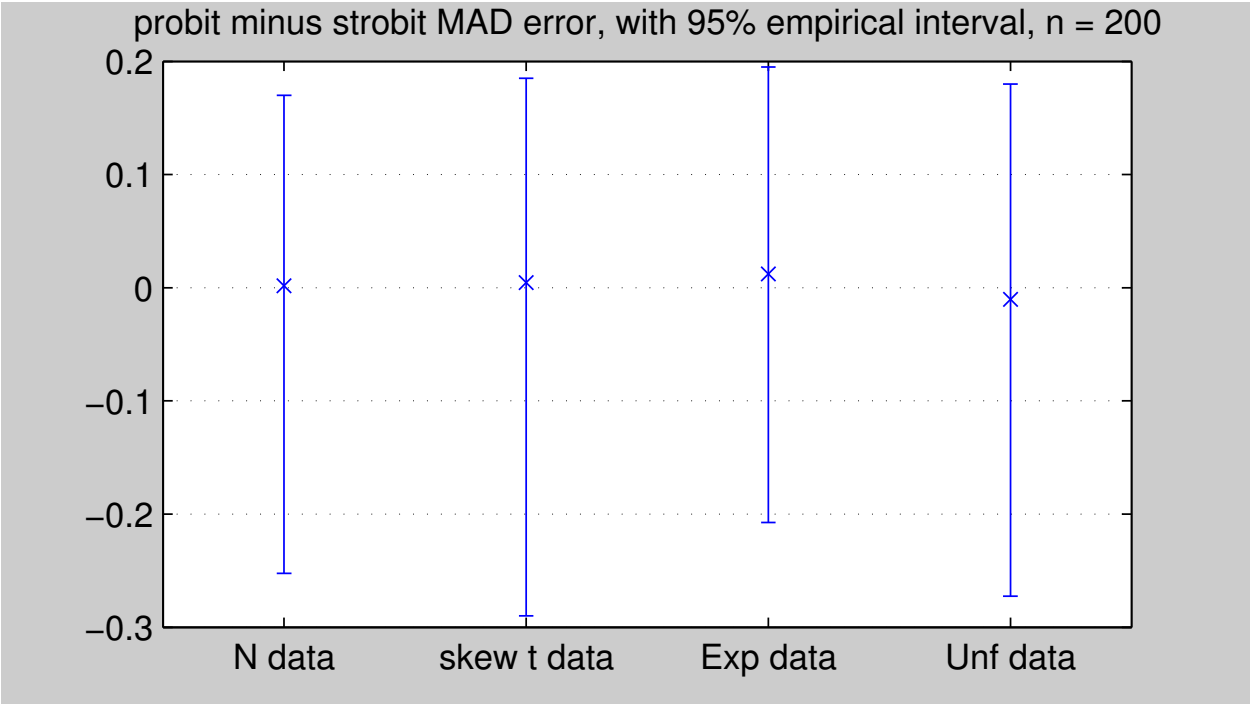
Scatterplot of probit minus strobit MAD error by negative sum of logs of outer category proportions for the three-category multiple simulation procedure.

Figure 16: Two-category mean MAD error difference with 95% interval, by data scenario, $n = 200$



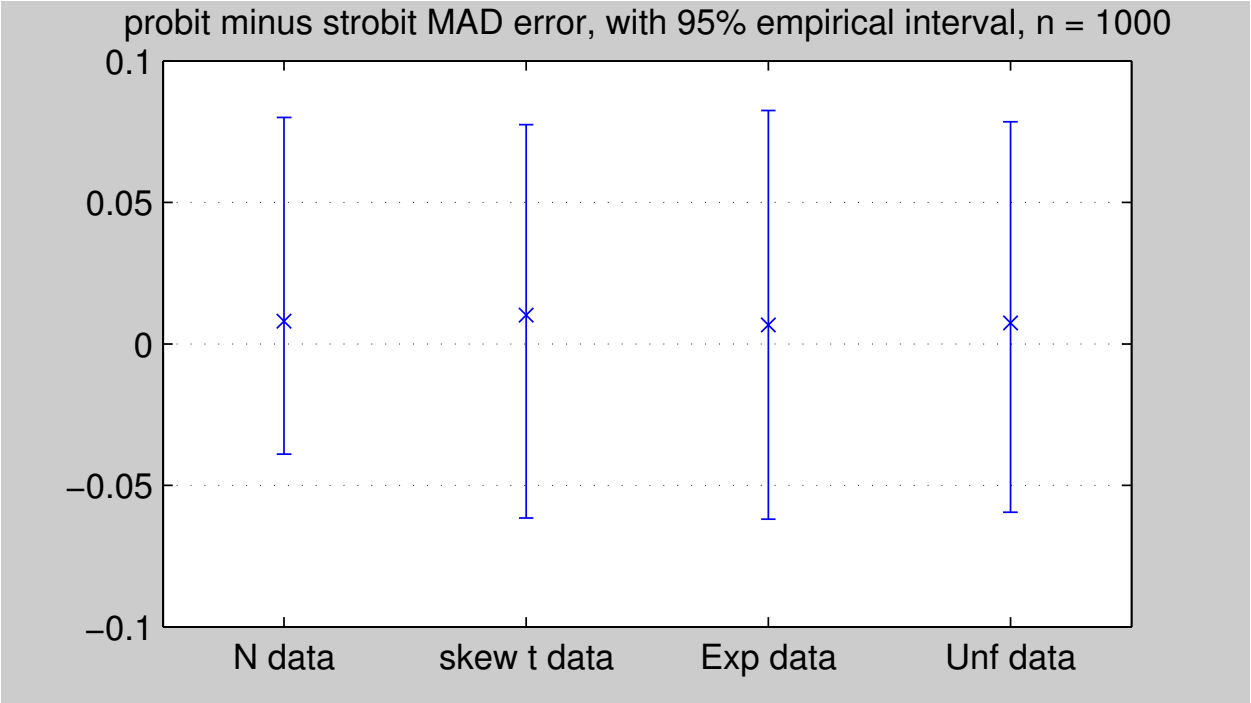
Mean probit minus strobit MAD error, with associated 95% empirical intervals, by data scenario, for the two-category multiple simulation procedure.

Figure 17: Three-category mean MAD error difference with 95% interval, by data scenario, $n = 200$



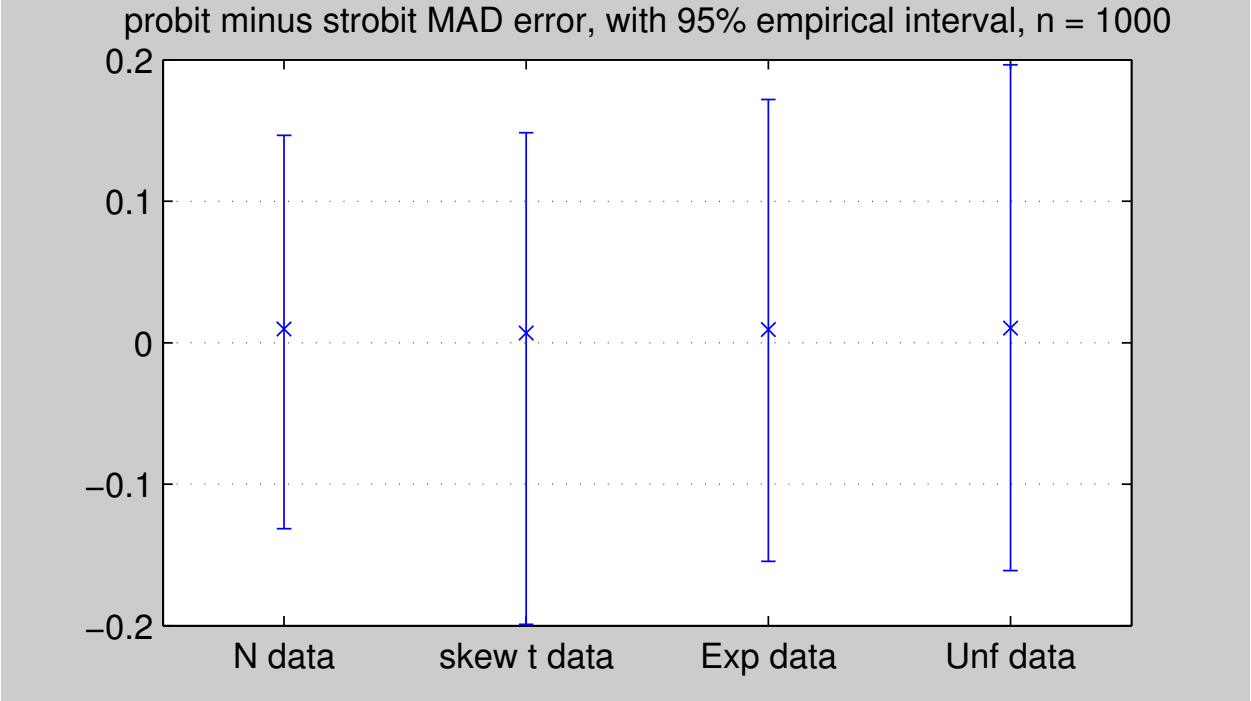
Mean probit minus strobit MAD error, with associated 95% empirical intervals, by data scenario, for the three-category multiple simulation procedure.

Figure 18: Two-category mean MAD error difference with 95% interval, by data scenario, $n = 1000$



Mean probit minus strobit MAD error, with associated 95% empirical intervals, by data scenario, for the two-category multiple simulation procedure.

Figure 19: Three-category mean MAD error difference with 95% interval, by data scenario, $n = 1000$



Mean probit minus strobit MAD error, with associated 95% empirical intervals, by data scenario, for the three-category multiple simulation procedure.