Statistics in Biosciences (2022) 14:533–549
https://doi.org/10.1007/s12561-022-09338-6

Fast Multivariate Probit Estimation via a Two-Stage Composite Likelihood

Bryan Ting[1] · **Fred Wright**[1,2] · **Yi-Hui Zhou**[1,2,3]

Received: 1 June 2021 / Revised: 5 January 2022 / Accepted: 9 January 2022 /
Published online: 15 February 2022

Abstract
The multivariate probit is popular for modeling correlated binary data, with an attractive balance of flexibility and simplicity. However, considerable challenges remain in computation and in devising a clear statistical framework. Interest in the multivariate probit has increased in recent years. Current applications include genomics and precision medicine, where simultaneous modeling of multiple traits may be of interest, and computational efficiency is an important consideration. We propose a fast method for multivariate probit estimation via a two-stage composite likelihood. We explore computational and statistical efficiency, and note that the approach sets the stage for extensions beyond the purely binary setting.

Keywords Composite · Likelihood · Multivariate · Probit · Two-stage · Two-step

1 Introduction

The modeling of binary data has a long history in biostatistics and biological sciences, with applications in areas as diverse as epidemiology [1] and expression quantitative trait analysis [2]. Despite the simplicity in outcomes, modeling binary data are not computationally trivial, often involving latent structures. Thus, the development of fast and effective modeling is of considerable importance.

The multivariate probit is a standard model for modeling correlated binary data, with advantages due to its flexibility in handling correlation structures and

✉ Yi-Hui Zhou
yihui_zhou@ncsu.edu

[1] Bioinformatics Research Center, North Carolina State University, 1 Lampe Drive, Raleigh, NC 27695, USA

[2] Statistics Department, North Carolina State University, 2311 Stinson Drive, Raleigh, NC 27695, USA

[3] Biological Sciences Department, North Carolina State University, 112 Derieux Place, Raleigh, NC 27695, USA

interpretability of parameters [3]. The approach is conceptually simple, in the sense that the underlying multivariate latent normal requires specification of only means and covariances. However, for likelihood parameter estimation, the integrals for calculating the likelihood from the multivariate cumulative normal distribution are computationally intensive [3, 4], e.g., as detailed in documentation for software such as the R package `mvProbit` [5].

We consider the standard multivariate probit, where binary components of the multivariate response $\boldsymbol{Y}$ are modeled as the binarized result of a latent multivariate distribution. For identifiability, we assume unit marginal latent variances, i.e., the covariance matrix is a correlation matrix. With K binary response components, this implies $\binom{K}{2}$ correlation values. We consider the N by P design matrix $\boldsymbol{X}$ as shared across components, as well as the $P \times K$ coefficient matrix $\boldsymbol{B}$, where N, P, and K are the number of observations, predictors, and components, respectively. The role of $\boldsymbol{X}$ is to serve as a predictor matrix in a regression framework for the latent outcome. For the multivariate probit, the role of the coefficients (in conjunction with the design matrix) can be viewed as specifying the mean for the latent multivariate normal probability, with the region of integration being $(-\infty, 0]$ or $(0, \infty)$ for a given component depending on whether the response is 0 or 1. Commonly, as we will do here, the mean is fixed at $\mathbf{0}$ and instead the coefficients help determine the boundaries of integration—a numerically equivalent representation.

Thus, for multivariate binary response $\boldsymbol{Y}$ with K components the full likelihood for observation i is:

$$L_{full}(\boldsymbol{\theta};\boldsymbol{y_i}) = \int_{A_{i1}} \dots \int_{A_{iK}} \phi(\boldsymbol{z_i}, \mathbf{0}, \boldsymbol{\Sigma}) d\boldsymbol{z_i}$$
$$\begin{cases} A_{ik} = (-\infty, \boldsymbol{x_i}\boldsymbol{\beta_k}] & y_{ik} = 1 \\ A_{ik} = (\boldsymbol{x_i}\boldsymbol{\beta_k}, \infty) & y_{ik} = 0 \end{cases},$$

where

$$\boldsymbol{\theta} = \{\boldsymbol{B}, \boldsymbol{\Sigma}\}$$

i corresponds to a given observation, and k a given component of $\boldsymbol{Y}$. The i, jth term in $\boldsymbol{\Sigma}$ represents the latent correlation between the ith and jth components of the response. The latent multivariate normal variable is assumed to have a mean vector of $\mathbf{0}$, with the constants of integration determined by whether observed values of the multivariate binary response are 0 or 1.

We do not apply constraints to the correlations, and issues of positive definiteness are addressed below. With PK coefficient parameters and $K(K-1)/2$ correlation parameters, the number of parameters grows quickly with increasing number of components.

Moffa and Kuipers [4] proposed a sequential expectation-maximization Monte Carlo method to estimate parameters in the multivariate probit. The approach builds on [3] and utilizes the truncated multivariate T distribution, with heavier tails than the normal. However, the approach can be computationally intensive, with variability in results due to the stochastic sampling.

Mullahy [6] proposed that multivariate probit estimation be performed via "chained" bivariate probits. Each element in the correlation matrix is estimated pairwise for components in the response, and coefficient estimates are obtained by averaging over coefficient estimates obtained from the bivariate pairings. The approach is computationally attractive, but statistical efficiency and other properties remain unclear. The chained bivariate probit approach is implemented in Mata's `bvpmvp()`, as a faster alternative to Stata's `mvprobit` [6]. Stata's `mvprobit` [7] and R's `mvProbit` both use the GHK (Geweke, Hajivassiliou and Keane) approach to simulate multivariate normal probabilities, and both can be computationally inefficient.

Fieuws and Verbeke [8] proposed taking a pairwise approach for estimating multivariate mixed models, given the computational challenges that can result when the number of components and/or the number of random effects is high. In this approach, $\frac{(K)(K-1)}{2}$ likelihoods are maximized separately instead of the full likelihood, where K is the number of components. In each of these likelihoods, only parameters associated with each pair of components are estimated. A given parameter can appear more than once across these likelihoods, in which case the average is taken over these estimates. The authors performed a simulation study using 1,000 replications of a trivariate mixed model with linear responses. They reported that the pairwise approach showed little bias in the estimates of the fixed effects, and showed little loss of efficiency for most of the parameters overall relative to the full trivariate approach. Fieuws et al. [9] applied this pairwise approach to a logistic model with random intercepts. Interestingly, the chained bivariate probit approach for the multivariate probit from [6] can be viewed as a special case of this pairwise approach.

Feddag [10] suggested using a composite pairwise likelihood approach in the context of estimating multivariate probit longitudinal models. In this formulation, an unconstrained covariance matrix is used (instead of correlation), and identifiability is assured by including a mean term and constraining coefficients to sum up to 0 across components. Simulations were performed with response variables of 3, 5, and 8 components with 50, 100, and 300 observations. Feddag [10] noted empirically that the general pairwise likelihood results retained nearly full statistical efficiency compared to using the full likelihood, but was much faster computationally. The standard deviations of coefficient estimates across simulations were similar between the full likelihood and their composite likelihood. For an example with 3 components and 300 observations, maximizing the pairwise likelihood took 0.16 minutes for a desktop computer to converge, whereas the full likelihood took 27.3 minutes.

Jin [11] also found good performance of a composite pairwise likelihood for binary data using a different model. Pairwise likelihood also performed well in terms of both efficiency and computation time. In a larger exploration of a multivariate normal model [12], included a similar two-stage composite likelihood for multivariate probit in simulations, where the first stage consisted of univariate marginals, and in the second stage bivariate marginals. However, they were chiefly concerned with analysis of data in familial units. Thus, their simulations were performed using models in which there were two mutual coefficients $\{\beta_0, \beta_1\}$ across components (that is, coefficients across components were constrained to equality) and where correlation parameters across components were either one ρ_1 or ρ_2, corresponding to

parent-offspring or sibling-sibling correlations. Each of their simulations used 2,000 families. The authors noted that the two-stage composite likelihood is faster to compute than the pairwise composite likelihood, and both are more computationally efficient than full maximum likelihood.

Ghosh [13] introduced a bivariate logistic model that includes an intermediate latent probit step. The approach, originally designed to handle bias in outcome-dependent sampling situation, has considerable flexibility in handling nuisance covariates. However, it is not easily extensible beyond $K = 2$.

To address issues of computational efficiency while retaining a balance of simplicity and flexibility, we introduce a novel two-stage composite likelihood approach for multivariate probit estimation. This approach is designed to be fast, and thus amenable to situations where many potential predictors are screened, such as with genome-wide association studies. Coefficient standard errors are obtained using a sandwich estimator appropriate for a composite likelihood. In contrast to [12], we focus upon multivariate probit models with unconstrained parameters, and show that our model can achieve impressive gains in computation time while largely maintaining statistical efficiency.

2 Methods

2.1 Two-stage Estimation

Two-stage, or "two-step," likelihood estimation [14] can be an option for analytically or computationally difficult likelihood and/or log-likelihood functions. In two-stage estimation, the original model is essentially split into two models, with the first embedded in the second. The first stage estimates parameters associated with the first likelihood, and the second stage the second likelihood. Following [14], suppose we start with a full log-likelihood and n independent observations:

$$\ln L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} \ln f(\boldsymbol{y}_{1i}, \boldsymbol{y}_{2i} \mid \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

The parameter vector $\boldsymbol{\theta}_1$ is associated via likelihood f_1 with data $\boldsymbol{x}_1, \boldsymbol{y}_1$, and in the first stage, the parameters in $\boldsymbol{\theta}_1$ are estimated by maximizing:

$$\ln L(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \ln f_1(\boldsymbol{y}_{1i} \mid \boldsymbol{x}_{1i}, \boldsymbol{\theta}_1).$$

In the second stage, the estimates of $\boldsymbol{\theta}_1$ from the first stage can be used as fixed inputs to maximize the conditional likelihood via f_2:

$$\ln L\left(\boldsymbol{\theta}_2 \mid \hat{\boldsymbol{\theta}}_{15}\right) = \sum_{i=1}^{n} \ln f_2\left(\boldsymbol{y}_{2i} \mid \boldsymbol{x}_{2i}, \boldsymbol{\theta}_2, \left(\boldsymbol{x}_{1i}, \hat{\boldsymbol{\theta}}_1\right)\right),$$

where $\boldsymbol{y}_{2i}$ is a subset of responses from the i'th observation of response $\boldsymbol{y}$, and $\boldsymbol{y}_{1i}$ is another subset. $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ are their counterparts in the design matrix. $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ partition the full parameter vector $\boldsymbol{\theta}$. Either or both stages can be considered misspecified likelihoods, and for certain problems, maximizing the conditional likelihood is equivalent to maximizing the full likelihood. Since this is not always true in general, the suitability of two-stage estimation for a particular model and dataset would warrant further consideration. At the very least, the asymptotic covariance matrix of the second stage parameters would need to account for the first stage parameter estimates being treated as fixed in second stage estimation [14–16], as we do here.

2.2 Composite Likelihood

A composite likelihood is formed by the product of so-called "associated" or "sub"-likelihoods that are individually proper likelihoods. Like two-stage estimation, composite likelihoods are a popular alternative when maximizing the full likelihood is computationally difficult.

Lindsay et al. [17] provide an overview of theoretical properties and construction strategies. For proper sub-likelihoods, the composite likelihood is generally consistent, but may suffer a loss in efficiency compared to the full likelihood [17]. Suppose there are A associated likelihoods (each of which in general involves all observations). Following [18], we write the composite likelihood:

$$L_{\text{comp}}(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{a=1}^{A} f(\boldsymbol{y};\boldsymbol{\theta})^{w_a},$$

where $\boldsymbol{\theta}$ is the parameter vector and w_a denotes a non-negative weight for the a'th associated likelihood. This weight parameter is fixed in advance, and is often 1 for all sub-likelihoods (in which case the parameter can be ignored). A higher weight affords greater import to a given associated likelihood, vice versa for lower weights.

The simplest composite likelihood is the unweighted independence likelihood, where each associated likelihood k corresponds to the k'th component:

$$L_{\text{ind}}(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{k=1}^{K} f(y_k;\boldsymbol{\theta}).$$

If dependence parameters are involved, a natural extension is the unweighted pairwise likelihood:

$$L_{\text{pair}}(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{i=1}^{K-1} \prod_{j=i+1}^{K} f(y_i, y_j;\boldsymbol{\theta}).$$

As discussed by [18], choice of weight parameters can depend upon the problem at hand. One example is clustered multivariate normal data. For pairwise likelihood, [19] recommended a weighting scheme of $1/((m_i - 1)(1 + 0.5(m_i - 1)))$, observations being $y_{ir}, r = 1, \ldots m_i$, where observations within the i'th cluster are correlated. This is a compromise between the $1/(m_i - 1)$ recommended by [12, 20, 21], which

is suitable for low dependencies-and weights of $1/(m_i(m_i-1))$, which is better for stronger dependencies. Another example is longitudinal data, where commonly proposed are weighting schemes that downweight pairs farther apart in time [18]. In fact, [19] found that using only adjacent pairs for the pairwise likelihood to be preferable over using all pairs.

Varin et al. [18] provide a review of composite likelihood methods and remarked upon their practical high statistical efficiency. However, the efficiency and asymptotic properties can depend importantly on specific of the full likelihood and the composite set-up, e.g., marginal versus conditional likelihoods and the complexity of the sub-likelihoods [18, 22].

2.3 Two-stage Composite Likelihood

We estimate parameters for the multivariate probit likelihood using a composite likelihood, and divide the composite likelihood estimation process into two stages. For both stages, we use an implied weight parameter of 1 for all associated likelihoods-in the interest of simplicity-as is often done [18].

In the first stage, we obtain coefficient estimates from a composite likelihood consisting of univariate marginals. Each associated likelihood involves one component from the response, which for our setting is the univariate probit:

$$\ln L_{uni}(\boldsymbol{B};\boldsymbol{y_i},\boldsymbol{x_i}) = \sum_{k=1}^{K} \ln f(y_{ik},\boldsymbol{x_i};\boldsymbol{\beta_k}).$$

As no parameters are shared across sub-likelihoods here, for estimation we can write:

$$\begin{aligned} max_{\boldsymbol{B}} \sum_{i=1}^{N} \ln L_{uni}(\boldsymbol{B};\boldsymbol{y_i},\boldsymbol{x_i}) &= max_{\boldsymbol{B}} \sum_{i=1}^{N}\sum_{k=1}^{K} \ln f(y_{ik},\boldsymbol{x_i};\boldsymbol{\beta_k}) \\ &= max_{\boldsymbol{\beta_1}} \sum_{i=1}^{N} \ln f(y_{i1},\boldsymbol{x_i};\boldsymbol{\beta_1}) + \cdots + max_{\boldsymbol{\beta_K}} \sum_{i=1}^{N} \ln f(y_{iK},\boldsymbol{x_i};\boldsymbol{\beta_k}). \end{aligned}$$

Since these associated likelihoods can be estimated independently, this simplifies the computational process. For example, R's `glm()` can provide coefficient estimates for each of the components.

In the second stage, we estimate the correlation parameters, using as inputs the coefficient estimates from the first stage. This plug-in approach can be justified from the fact that the first stage did not include the correlations, as well as consistency arguments that apply to each of the sub-likelihoods. Here each associated likelihood involves a pair of components, each a bivariate probit:

$$\ln L_{pair}(\boldsymbol{\Sigma};\boldsymbol{y_i},\boldsymbol{x_i},\hat{\boldsymbol{B}}) = \sum_{j=1}^{K-1}\sum_{k=j+1}^{K} \ln f(y_{ij},y_{ik},\boldsymbol{x_i};\rho_{jk},\hat{\boldsymbol{\beta}}_{\boldsymbol{j}},\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}).$$

Again, no parameters are shared across associated likelihoods, so we approach maximization component-wise:

$$\begin{aligned} max_{\boldsymbol{\Sigma}} \sum_{i=1}^{N} \ln L_{pair}(\boldsymbol{\Sigma}; \boldsymbol{y_i}, \boldsymbol{x_i}, \hat{\boldsymbol{B}}) &= max_{\boldsymbol{\Sigma}} \sum_{i=1}^{N} \sum_{j=1}^{K-1} \sum_{k=j+1}^{K} \ln f(y_{ij}, y_{ik}, \boldsymbol{x_i}; \rho_{jk}, \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{k}}) \\ &= \Big(max_{\rho_{1,2}} \sum_{i=1}^{N} \ln f(y_{i1}, y_{i2}, \boldsymbol{x_i}; \rho_{1,2}, , \hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}) + \cdots \\ &+ max_{\rho_{K-1,K}} \sum_{i=1}^{N} \ln f(y_{i(K-1)}, y_{iK}, \boldsymbol{x_i}; \rho_{K-1,K}, \hat{\boldsymbol{\beta}}_{\boldsymbol{K-1}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{K}}) \Big). \end{aligned}$$

The primary gain in computational efficiency arises from this component-wise estimation, which we can implement using simple maximization routines such as R's `optim()`. Functions such as `pbivnorm()` [23] can provide fast bivariate probit probability calculations.

In contrast to models where parameters are constrained, such as the familial models explored by [12], here maximization can proceed independently. If the overall estimated K by K correlation matrix turns out not to be positive semidefinite, routines such as R's `nearPD()` [24] can find the nearest positive definite correlation matrix as a post-processing step after estimation. However, for most real-world applications, especially the applications we have in mind-where sample sizes are large relative to the number of components and predictors-such a step is unlikely to be needed. For example, a genomics dataset that we use below for simulations has 728 observations.

Maximization via the two-stage modeling procedure can conceivably produce a non-positive definite correlation matrix estimate, which is somewhat analogous to difficulties in variance estimation for variance component models [25]. We first note that, provided the correlation matrix is strictly positive definite (the case of interest), the large n consistency of composite likelihood ensures that the composite maximum likelihood estimate will be positive definite, with probability approaching one. This result follows from a continuous mapping theorem argument. In practice, all of our simulations and real data examples resulted in positive definite correlation estimates, except for a small proportion (4–8 components) of simulations for small sample sizes of 100. In these rare instances, we apply the method based on nearest Frobenius norm [26].

Composite likelihoods can be considered misspecified likelihoods, as the sub-likelihoods do not fully reflect data dependencies. [14] describes a robust variance estimator to account for both misspecification and the two-stage nature of the estimation process—essentially a "sandwich" version of the Murphy–Topel variance estimator [16] for two-stage models. Following [14], let $\boldsymbol{V_S}(\boldsymbol{\theta_1})$ denote the robust variance estimator for $\hat{\boldsymbol{\theta}}_1$ estimated in the first stage, and $\boldsymbol{V_S}(\boldsymbol{\theta_2})$ for $\hat{\boldsymbol{\theta}}_2$ in the second stage, with $\boldsymbol{Cov_S}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ the covariance between them. We have

$$\begin{aligned} \boldsymbol{V_S}(\boldsymbol{\theta_1}) &= \boldsymbol{V_1 V_1^{*-1} V_1} = \boldsymbol{V_{S1}} \\ \boldsymbol{Cov_S}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) &= \boldsymbol{V_1 R^T V_2} - \boldsymbol{V_{S1} C^{*T} V_2} \\ \boldsymbol{V_S}(\boldsymbol{\theta_2}) &= \boldsymbol{V_2 V_2^{*-1} V_2} + \boldsymbol{V_2}(\boldsymbol{C^* V_{S1} C^{*T}} - \boldsymbol{R V_1 C^{*T}} - \boldsymbol{C^* V_1 R^T})\boldsymbol{V_2} \\ &= \boldsymbol{V_{S2}} + \boldsymbol{V_2}(\boldsymbol{C^* V_{s1} C^{*T}} - \boldsymbol{R V_1 C^{*T}} - \boldsymbol{C^* V_1 R^T})\boldsymbol{V_2}, \end{aligned}$$

where $\boldsymbol{V_1}$ is the non-robust (naive) likelihood-based asymptotic variance estimator for the stage one parameters $\boldsymbol{\theta_1}$ based upon the stage one log-likelihood $\ln L_1(\boldsymbol{\theta_1})$

(i.e., the expected value of the negative second derivatives), and $\boldsymbol{V}_{\mathbf{1}}^{*}$ the expected value of the matrix of outer product of gradients. Similarly, $\boldsymbol{V}_{\mathbf{2}}$ is the non-robust asymptotic variance estimator for the stage two parameters $\boldsymbol{\theta}_{\mathbf{2}}$ based upon the stage two conditional log-likelihood $\ln L_2(\boldsymbol{\theta}_{\mathbf{2}} \mid \boldsymbol{\theta}_{\mathbf{1}})$, and $\boldsymbol{V}_{\mathbf{2}}^{*}$ the expected value of the matrix of outer gradients. $\boldsymbol{C}^{*}$ is the sub-matrix of the expected value of the negative second derivatives based on $\ln L_2(\boldsymbol{\theta}_{\mathbf{2}} \mid \boldsymbol{\theta}_{\mathbf{1}})$, the rows corresponding to $\boldsymbol{\theta}_{\mathbf{2}}$ and the columns corresponding to $\boldsymbol{\theta}_{\mathbf{1}}$. $\boldsymbol{R}$ is the sub-matrix of the expected value of the negative second derivatives based on $\ln L_2(\boldsymbol{\theta}_{\mathbf{2}} \mid \boldsymbol{\theta}_{\mathbf{1}})$ and $\ln L_1(\boldsymbol{\theta}_{\mathbf{1}})$, the rows corresponding to $\boldsymbol{\theta}_{\mathbf{2}}$ and the columns corresponding to $\boldsymbol{\theta}_{\mathbf{1}}$. $\boldsymbol{V}_{s\mathbf{1}}$ is the usual sandwich estimator for the stage one parameters, and $\boldsymbol{V}_{s\mathbf{2}}$ that of the second stage parameters (treating stage one parameters as fixed). Empirical plug-in estimates for the matrix elements are obtained by taking the mean across observations (using the final parameter estimates as inputs). Once estimated, these matrices can be used to calculate the robust Murphy–Topel estimate of variance [16].

3 Examples

3.1 Six Cities

The *Six Cities* dataset has been a popular choice for comparing multivariate probit estimation methodologies. We performed our two-stage composite likelihood estimation upon this dataset, and compared it to the results of [3, 4]. We were chiefly concerned with run-time and statistical efficiency, as judged by coefficient standard errors.

In the *Six Cities* data, wheezing status at ages 7, 8, 9, and 10 for 537 children were recorded as 0 or 1 to serve as the multivariate response, for four components with binary observations. Coefficients (shared across all components) included the intercept, age centered at 9, maternal smoking status (1 or 0), and an interaction variable between maternal smoking status and age. These were represented by β_0, β_1, β_2 and β_3, respectively. Note that the four components share coefficients, i.e., $\boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \boldsymbol{\beta_3} = \boldsymbol{\beta_4} = \{\beta_0, \beta_1, \beta_2, \beta_3\}$. The covariance (correlation) matrix has 6 off-diagonal entries, corresponding to correlations between wheezing status between various ages. Standard errors were calculated using the robust approach described in the previous section. 250 bootstrapped replications and estimates were also performed for comparison.

Using our two-stage composite likelihood, the *Six Cities* coefficient and correlation estimates are very similar to those previously published by [3, 4]. Parameter estimation took about $\frac{1}{40}$ of a second for our model on a Windows 2.70 GHz Intel i7-7500 laptop. The bootstrapped standard errors and empirical standard errors obtained from the original data are similar to each other, and also similar to those provided by [3, 4]. In summary, for these data the estimates do not reflect apparent loss in statistical efficiency, and the correspondence with bootstrapped standard errors indicates appropriateness of the robust variance estimates (Table 1).

Interestingly, the two-stage composite likelihood produces estimates that achieved a higher log-likelihood when inputted into the full information likelihood than did

Table 1 Comparison of *Six Cities* mean parameter estimates between [3, 4], and the Two-Stage Composite Likelihood

Six Cities Estimation Comparisons							
Param.	Chib and Greenberg (1998)		Moffa and Kuipers (2014)		Two-stage CL (2019)		
	Est.	SE	Est.	SE	Est.	BSE	ESE
β_0	− 111.8	(6.5)	− 112.3	(6.2)	− 112.6	(6.5)	(6.3)
β_1	− 7.9	(3.3)	− 7.9	(3.1)	− 7.7	(3.1)	(3.1)
β_2	15.2	(10.2)	15.9	(10.1)	17.1	(10.6)	(10.1)
β_3	3.9	(5.2)	3.8	(5.1)	3.7	(4.9)	(4.9)
σ_{12}	58.4	(6.8)	58.3	(6.6)	59.1	(6.5)	(6.6)
σ_{13}	52.1	(7.6)	52.2	(7.1)	53.1	(7.3)	(7.2)
σ_{14}	58.6	(9.5)	57.8	(7.4)	59.1	(7.5)	(7.2)
σ_{23}	68.8	(5.1)	68.6	(5.6)	69.2	(5.7)	(5.6)
σ_{24}	56.2	(7.7)	55.8	(7.4)	57.5	(7.5)	(7.3)
σ_{34}	63.1	(7.7)	62.7	(6.7)	64.1	(6.4)	(6.6)
Est. Log-Lik	− 794.94	(0.69)	− 794.95	(0.82)	− 794.76	(–)	(0.00)
Corr. Log-Lik	− 794.70		− 794.61		− 794.76		

250 replications were done for the Two-Stage Composite Likelihood to calculate mean parameter estimates and the mean log-likelihood value (its empirical standard error), and 250 bootstrapped replications for the bootstrapped standard errors. Empirical standard errors were calculated using the robust Murphy–Topel variance estimator. Parameter estimation took about 0.025 s for the Two-Stage Composite Likelihood

the log-likelihoods from [3] or [4]. However, as pointed out by [4], the stochastic nature of their processes (leading to noticeable variance across replications of log-likelihood estimations upon the same dataset) may reduce the log-likelihood. They thus supply a correction calculation for the log-likelihood. The two-stage composite likelihood does not require such a correction. Furthermore, for the two-stage composite likelihood, each replication and bootstrapped replication produced positive semidefinite estimates for the covariance matrix, so no post-processing steps were needed for positive definiteness.

3.2 MEPS

Started in 1996, The Medical Expenditure Panel Survey (MEPS) is a set of surveys containing data on how American families and individuals use health services [27]. The R package `GJRM` [28] provides a 2008 MEPS dataset with 18,592 observations of individual characteristics and their binary-coded disease statuses for diabetes, hyperlipidemia, and hypertension. The function `gjrm()` of `GJRM` can perform a variety of semi-parametric model estimations, including fully parametric univariate, bivariate, and trivariate probit estimations using a general penalized maximum likelihood approach in conjunction with smoothing set-up via penalized regression

splines [28]. Here, we use the MEPS dataset to compare `gjrm()` and the Two-Stage Composite Likelihood in estimating a trivariate probit. The trivariate response is the three disease statuses, and individual characteristics serve as predictors: body mass index (BMI), age (in years), sex (1 for male, 0 for female), education (in years), income (log-transformed), race (coded as white, black, Native American, or other), and region (northeast, midwest, south, or west). 18,273 observations were retained after excluding those with incomes listed as zero.

The mean run-time for `gjrm()` on a Windows 2.70 GHz Intel i7-7500 laptop was about 32 s, whereas it was about 12 s for the Two-Stage Composite Likelihood. The coefficient estimates for the three components of Diabetes, Hyperlipidemia, and Hypertension are displayed below in Table 2. The correlation parameter estimates are included at the bottom. Like with the *Six Cities* example, no post-processing for positive definiteness was needed.

The parameter estimates were generally similar, and the bootstrapped standard errors for the two methods were nearly identical for both coefficients and correlation parameters. The reported standard errors from `gjrm()` and the empirical standard errors from the Two-Stage CL were generally close for coefficients. However, the reported standard errors from `gjrm()` were noticeably higher than the `gjrm()` bootstrapped standard errors, as well as both the empirical and bootstrapped standard errors from the Two-Stage CL (Table 2).

4 Simulations

4.1 Run-time Comparisons with the Chained Bivariate Probit

Simulations were performed following the set-up described by [6]. That article found that the chained bivariate probit approach was much faster than a simulation-based maximum likelihood approach. Following [6], the number of components considered was 4 and 8, the number of predictors considered was 5 and 9 (including intercepts), and the number of observations varied among 2000, 5000, and 10,000. We also included additional observation counts of 100 and 500. 500 simulations were ran for each of these 20 combinations for each approach.

For coefficients, the intercepts ranged step-wise from 0.7 to $-$ 0.7 from the first component to the eighth, and the coefficients on the predictor of interest alternated between the values of 0.0 and 0.5. Other coefficient parameters were set to 0.0. When only four components were needed for a given combination of settings, the first four components were used. For our analysis, design matrices were constructed using an intercept and the first 3 or 7 principal component values from a genetic dataset consisting of ternary data obtained by subsampling a random set of genetic markers from HapMap data [29], with one arbitrarily chosen of the genetic markers (the first one) as the predictor of interest.

For each combination of number of observations, number of components, and number of predictors, a design matrix was sampled (with replacement) from the original 728 observations. The multivariate response was then simulated anew for each of the 500 simulations per combination. Run-time comparisons were made

Table 2 Comparison of estimates produced by `gjrm()` and the Two-Stage Composite Likelihood, displayed by component and coefficient, along with the correlation parameter estimates

MEPS Estimation Comparisons							
Comp.	Param.	Gjrm()			Two-stage CL		
		Est.	SE	BSE	Est.	ESE	BSE
Diabetes	Intercept	− 375.4	20.6	21.6	− 360.9	21.0	22.0
	BMI	5.3	0.2	0.2	5.1	0.2	0.3
	Age	3.9	0.1	0.2	3.9	0.1	0.2
	Sex	4.9	3.1	3.1	4.5	3.1	3.1
	Education	− 3.7	0.5	0.5	− 4.1	0.5	0.5
	Income (10,000 s)	− 5.5	1.8	1.7	− 5.9	1.9	1.7
	Race (Black)	14.9	3.9	4.8	15.4	3.9	4.8
	Race (Nat. Amer.)	44.6	12.5	12.8	43.5	12.8	12.8
	Race (Other)	25.4	5.8	5.8	25.3	5.9	5.9
	Region (Midwest)	− 13.8	5.3	5.4	− 13.4	5.4	5.3
	Region (South)	− 3.1	4.5	4.4	− 3.0	4.6	4.4
	Region (West)	− 3.3	4.9	4.9	− 2.6	5.0	4.9
Hyperlipidemia	Intercept	− 400.3	15.3	14.6	− 400.4	15.3	14.7
	BMI	3.4	0.2	0.2	3.5	0.2	0.2
	Age	4.8	0.1	0.1	4.8	0.1	0.1
	Sex	13.5	2.2	2.4	14	2.2	2.4
	Education	0.9	0.4	0.4	0.8	0.4	0.4
	Income (10,000s)	1.2	1.3	1.3	1.2	1.3	1.3
	Race (Black)	− 13.0	3.1	3.4	− 12.7	3.0	3.4
	Race (Nat. Amer.)	13.1	11.0	10.7	13.2	11.3	10.7
	Race (Other)	15.6	4.1	4.5	15.7	4.1	4.5
	Region (Midwest)	− 7.9	3.8	3.6	− 8.2	3.8	3.6
	Region (South)	0.9	3.3	3.2	0.8	3.3	3.2
	Region (West)	− 7.3	3.6	3.8	− 7.3	3.6	3.8
Hypertension	Intercept	− 351.1	15.1	15.6	− 350	14.8	15.6
	BMI	5.7	0.2	0.2	5.7	0.2	0.2
	Age	4.8	0.1	0.1	4.8	0.1	0.1
	Sex	11.7	2.3	2.3	12.1	2.3	2.3
	Education	− 0.6	0.4	0.4	− 0.7	0.4	0.4
	Income (10,000s)	− 8.5	1.3	1.2	− 8.5	1.3	1.2
	Race (Black)	27.9	2.9	2.8	27.8	2.9	2.8
	Race (Nat. Amer.)	25.9	10.8	11.1	26.3	10.5	11.2
	Race (Other)	15.1	4.3	4.1	15.0	4.4	4.1
	Region (Midwest)	− 7.2	3.9	4.0	− 7.5	3.9	4.1
	Region (South)	3.1	3.4	3.2	2.9	3.3	3.3
	Region (West)	− 9.2	3.7	3.5	− 9.3	3.7	3.5
Diabetes	Hyperlipidemia	0.41	0.04	0.02	0.41	0.02	0.02
Diabetes	Hypertension	0.35	0.03	0.02	0.35	0.02	0.02
Hyperlipidemia	Hypertension	0.41	0.03	0.02	0.41	0.01	0.02

The standard errors from `gjrm()` come from its native reporting; empirical standard errors for the Two-Stage Composite Likelihood were calculated using the robust Murphy–Topel variance estimator. 250 replications were performed to record the bootstrapped standard errors; parameter estimation took, on average, about 32 s for `gjrm()`, and 12 s for the Two-Stage Composite Likelihood. Coefficient estimates multiplied by 100; correlation estimates displayed without rounding

versus reported values from [6], in which the previous author used a desktop computer with a higher clock speed (iMac 3.4 GHz Intel Core i7) than used here. The number of times the post-processing step was invoked for the two-stage composite likelihood to ensure positive definiteness is shown, as well (Table 3).

For each combination of settings, the two-stage composite likelihood was the fastest. [6]'s reported results were faster than the re-coded version of the chained bivariate probit, possibly at least partly due to processor specifications. As also observed by [6], the number of components had the greatest effect on run-time, followed by number of observations, and the number of predictors had the least effect. In fact, for the two-stage composite likelihood, having nine predictors instead of five did not appear to result in longer run-times.

When the observation counts were low, on occasion one of the bivariate probits in the re-coded chained bivariate probit would fail to complete. When this occurred, the particular simulation was skipped over for the re-coded chained bivariate probit and excluded from the run-time average. However, this only occurred seven and five times out of 500 observations, respectively, for five and nine predictors when the number of components was eight and the observation count was 100. The two-stage composite likelihood did not experience this issue.

As expected, the two-stage composite likelihood did not produce non-positive semidefinite estimates for the covariance matrix when the number of observations was sufficiently large. Non-positive semidefiniteness occurred frequently when the number of observations was 100 and the number of components was eight, but rarely when the number of components was four. The additional time to check for positive semidefiniteness and, if necessary, perform the post-processing step for positive definiteness was included in the run-time estimates for the two-stage composite likelihood.

4.2 Coverage Percentages for the Two-Stage Composite Likelihood

Using a similar set-up (and the same original data) as for the run-time comparisons, simulations were performed to gauge the 95% coverage percentages for the two-stage composite likelihood. The number of predictors was fixed at four, and the number of observations and components were varied for the simulations. 10,000 simulations of each combination of observations and number of components were ran. The 95% coverage percentages for the coefficient and correlation parameters are displayed below. Coverage is near the target 95% in all instances for coefficients, similarly for correlation parameters as observation counts approach 800. The post-processing step for positive definiteness was invoked only once, for when the number of observations was 200 and the number of components was five (Tables 4 and 5).

Table 3 Mean run-time comparisons between [6]'s `bvpmvp` & the Two-Stage Composite Likelihood, by combinations of number of observations, number of components, and number of predictors

Mean Run-time Comparisons						
# Obs	# Comp	# Pred	`bvpmvp`	Re-coded `bvpmvp`	Two-stage CL	PD Step
100	4	5		0.5	0.0	1
100	4	9		0.5	0.0	3
100	8	5		2.3	0.1	189
100	8	9		2.5	0.1	239
500	4	5		0.7	0.1	0
500	4	9		0.8	0.1	0
500	8	5		3.4	0.3	0
500	8	9		3.5	0.3	0
2000	4	5	1	1.7	0.3	0
2000	4	9	1	1.8	0.3	0
2000	8	5	5	7.9	1.2	0
2000	8	9	8	8.0	1.1	0
10,000	4	5	2	6.5	1.3	0
10,000	4	9	3	7.1	1.4	0
10,000	8	5	14	31.8	5.9	0
10,000	8	9	19	33.2	5.9	0
50,000	4	5	12	34.3	7.9	0
50,000	4	9	18	35.4	7.4	0
50,000	8	5	65	146.3	28.8	0
50,000	8	9	86	157.1	29.0	0

Number of times that the post-processing step for positive definiteness was invoked for the Two-Stage CL is shown, as well. Run-times for [6] are rounded to the nearest second as originally reported, nearest tenth for the re-coded chained bivariate probit and the Two-Stage CL. Each combination for the re-coded chained bivariate probit and the Two-Stage CL was simulated 500 times. [6]'s results were performed on an iMac 3.4 GHz Intel Core i7. The re-coded chained bivariate probit and the Two-Stage CL were ran on a Windows 2.70 GHz Intel i7-7500. The re-coded chained bivariate probit used R's `zelig()` [30, 31] from the `Zelig` package to perform the bivariate probits

5 Discussion

Our proposed two-stage composite likelihood for the multivariate probit produces results similar to previously published results, for both parameter estimations and standard errors. Bootstrap comparisons show that the robust variance estimates provide accurate standard errors. Furthermore, these standard errors are comparable to those of full likelihood or those of alternate methods, suggesting little loss in statistical efficiency.

Run-times for the two-stage composite likelihood compare favorably to the chained bivariate probit approach, which was already much faster than the approach using simulated maximum likelihood. The effects of increasing settings such as the number of observations, number of components, and number of

Table 4 95% Coverage percentages for the coefficient parameters of the two-stage composite Likelihood, using 10,000 simulations

Two-stage Composite Likelihood 95% Coefficient Parameter Coverage Percentages																				
Obs	Comp	$B_{1,1}$	$B_{1,2}$	$B_{1,3}$	$B_{1,4}$	$B_{2,1}$	$B_{2,2}$	$B_{2,3}$	$B_{2,4}$	$B_{3,1}$	$B_{3,2}$	$B_{3,3}$	$B_{3,4}$	$B_{4,1}$	$B_{4,2}$	$B_{4,3}$	$B_{4,4}$	$B_{5,1}$	$B_{5,2}$	$B_{5,3}$	$B_{5,4}$
200	3	94.8	95.0	95.0	95.3	94.9	94.7	95.0	94.8	95.3	95.1	94.6	94.5								
200	4	94.9	95.2	94.8	95.3	95.0	94.9	95.0	95.2	95.0	95.0	94.8	94.8	95.3	95.1	95.3	95.0				
200	5	95.0	94.7	95.0	95.1	95.3	95.2	95.1	95.2	95.3	94.9	94.5	94.9	95.1	95.0	94.6	95.2	94.9	94.7	94.9	94.7
400	3	95.6	95.2	94.9	95.4	95.2	94.9	95.0	95.0	95.0	95.2	95.2	95.2								
400	4	95.1	95.3	94.9	95.2	95.0	94.8	95.0	95.1	94.8	94.7	94.7	95.3	95.2	94.6	95.1	95.1				
400	5	95.2	95.4	94.9	95.3	95.0	95.4	94.8	95.2	95.2	95.0	94.8	94.9	94.7	95.0	95.4	95.0	94.8	94.9	94.8	95.3
800	3	94.9	94.8	95.1	94.9	95.1	94.9	94.8	94.8	94.5	94.9	94.9	94.9								
800	4	95.1	95.1	95.2	95.2	95.1	95.0	95.0	95.4	95.4	95.1	94.6	95.2	95.2	95.0	94.7	95.2				
800	5	94.9	95.3	95.2	95.1	95.0	94.8	95.1	95.7	95.3	95.3	94.8	95.0	94.9	95.2	94.5	94.9	94.8	95.1	94.8	95.0
1600	3	95.4	95.1	94.7	95.1	95.0	95.0	95.1	94.8	95.1	95.2	95.3	95.0								
1600	4	95.1	95.0	94.8	94.9	94.9	95.2	95.0	95.0	95.3	94.5	95.2	94.8	95.2	95.2	95.2	94.9				
1600	5	94.8	94.8	94.8	95.2	94.9	94.9	95.3	95.0	94.8	94.8	95.0	94.8	95.0	95.1	94.9	95.2	95.2	95.5	94.8	95.1
3200	3	94.9	95.0	94.9	94.7	94.9	95.2	95.0	94.9	95.1	95.4	94.9	95.0								
3200	4	94.8	95.1	94.7	95.0	94.9	95.3	94.8	95.3	95.0	95.0	95.2	94.8	95.0	95.4	95.0	95.1				
3200	5	95.3	95.4	95.0	95.3	95.4	95.3	95.0	95.0	94.8	95.0	94.7	94.8	94.6	95.0	95.1	94.9	95.3	94.9	94.8	95.0
6400	3	95.0	95.1	94.8	94.9	94.9	94.7	94.8	94.9	94.9	95.1	95.2	94.9								
6400	4	95.1	94.8	95.0	95.1	94.9	95.3	94.6	94.8	94.8	94.8	95.3	94.9	94.8	94.8	94.9	95.1				
6400	5	95.3	94.9	95.2	95.1	95.2	95.0	94.9	95.1	95.5	95.1	94.6	95.4	94.8	95.2	94.5	95.0	94.7	95.2	94.7	94.6

$B_{m,p}$ corresponds to the coefficient associated with the m'th component and p'th predictor

Table 5 95% Coverage Percentages for the correlation parameters of the Two-Stage Composite Likelihood, using 10,000 simulations

Two-stage Composite Likelihood 95% Correlation Parameter Coverage Percentages											
Obs	Comp	$\Sigma_{1,2}$	$\Sigma_{1,3}$	$\Sigma_{1,4}$	$\Sigma_{1,5}$	$\Sigma_{2,3}$	$\Sigma_{2,4}$	$\Sigma_{2,5}$	$\Sigma_{3,4}$	$\Sigma_{3,5}$	$\Sigma_{4,5}$
200	3	93.1	93.2			94.1					
200	4	93.1	93.2	93.5		93.9	93.5		94.4		
200	5	92.9	92.6	93.4		93.6	93.2	93.6	93.9	93.1	93.9
400	3	94.2	94.3			94.0					
400	4	94.2	93.9	93.8		94.5	94.5		94.3		
400	5	94.0	94.0	93.7	93.7	94.3	94.3	94.2	94.2	94.3	94.8
800	3	94.5	94.6			94.9					
800	4	94.9	94.7	94.5		94.3	94.6		94.8		
800	5	94.8	94.4	94.4	95.0	94.6	94.4	94.6	94.8	95.0	94.5
1600	3	94.8	95.2			94.7					
1600	4	94.9	94.6	95.3		94.7	95.0		94.8		
1600	5	95.0	94.4	94.4	95.3	94.4	94.8	94.8	95.0	95.1	94.9
3200	3	95.0	94.7			94.8					
3200	4	94.7	94.6	94.9		94.8	95.0		95.0		
3200	5	94.6	94.8	95.3	95.1	94.9	94.7	95.1	95.4	95.2	95.0
6400	3	95.0	95.0			94.8					
6400	4	94.7	94.8	95.0		94.9	95.0		94.4		
6400	5	95.0	94.8	95.0	94.9	94.8	95.2	94.8	95.0	95.4	94.7

The number of predictors was fixed at four, with the number of observations and components varied. $\Sigma_{k,l}$ corresponds to the correlation parameter associated with the k'th and l'th components

predictors has similar effects to that experienced by the chained bivariate probit. Under simulation, our approach produces near nominal confidence coverage.

A possible next step would be to extend this approach to incorporate heterogeneous multivariate responses, i.e., where the response can include both binary and continuous components. Such an approach would include bivariate normal densities for continuous-continuous pairs, as well as likelihoods for binary-continuous pairs. For binary-continuous pairs, the joint likelihood can be re-stated as the product of the marginal density of the continuous component multiplied against the conditional density of the binary component upon the continuous component.

Further considerations could include heteroskedasticity, i.e., non-constant variance across predictor values, which would further expand the range of applications.

Acknowledgements This work was supported in part by a Game Changing Research Initiative Grant from NC State, and grants from Cystic Fibrosis Foundation KNOWLE18XX0, KNOWLE21XX0.

References

1. Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG Jr, Speizer FE (1993) An association between air pollution and mortality in six US cities. N Engl J Med 329(24):1753–1759
2. Li G, Jima D, Wright FA, Nobel AB (2018) Ht-eqtl: integrative expression quantitative trait loci analysis in a large number of human tissues. BMC Bioinformatics 19(1):1–11
3. Chib S, Greenberg E (1998) Analysis of multivariate probit models. Biometrika 85(2):347–361
4. Moffa G, Kuipers J (2014) Sequential Monte Carlo em for multivariate probit models. Comput Stat Data Anal 72:252–272. https://doi.org/10.1016/j.csda.2013.10.019
5. Henningsen A (2019) "mvprobit". CRAN
6. Mullahy J (2016) Estimation of multivariate probit models via bivariate probit. Stand Genomic Sci 16(1):37–51
7. Cappellari L, Jenkins SP (2003) Multivariate probit regression using simulated maximum likelihood. Stand Genomic Sci 3(3):278–294. https://doi.org/10.1177/1536867X0300300305
8. Fieuws S, Verbeke G (2006) Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. Biometrics 62(2):424–431. https://doi.org/10.1111/j.1541-0420.2006.00507.x
9. Fieuws S, Verbeke G, Molenberghs G (2007) Random-effects models for multivariate repeated measures. Stat Methods Med Res 16(5):387–397. https://doi.org/10.1177/0962280206075305
10. Feddag M-L (2013) Composite likelihood estimation for multivariate probit latent traits models. Commun Stat Theory Methods 42(14):2551–2566. https://doi.org/10.1080/03610926.2010.538793
11. Jin Z (2009) On some aspects of composite likelihood. PhD dissertation, University of Toronto
12. Zhao Y, Joe H (2005) Composite likelihood estimation in multivariate data analysis. Can J Stat 33(3):335–356. https://doi.org/10.1002/cjs.5540330303
13. Ghosh A, Wright FA, Zou F (2013) Unified analysis of secondary traits in case–control association studies. J Am Stat Assoc 108(502):566–576. https://doi.org/10.1080/01621459.2013.793121
14. Hardin JW (2002) The robust variance estimator for two-stage models. Stand Genomic Sci 2(3):253–266. https://doi.org/10.1177/1536867X0200200302
15. Greene WH (2002) Econometric analysis, 5th edn. Pearson Education, Pearson
16. Murphy KM, Topel RH (1985) Estimation and inference in two-step econometric models. J Bus Econ Stat 3(4):370–379
17. Lindsay B, Yi G, Sun J (2011) Issues and strategies in the selection of composite likelihoods. Stat Sin 21:71–105
18. Varin C, Reid N, Firth D (2011) An overview of composite likelihood methods. Stat Sin 21(1):5–42
19. Joe H, Lee Y (2009) On weighting of bivariate margins in pairwise likelihood. J Multivar Anal 100(4):670–685. https://doi.org/10.1016/j.jmva.2008.07.004
20. Kuk A, Nott D (2000) A pairwise likelihood approach to analyzing correlated binary data. Stat Probab Lett 47:329–335. https://doi.org/10.1016/S0167-7152(99)00174-1
21. LeCessie S, van Houwelingen JC (1994) Logistic regression for correlated binary data. Appl Stat 43:95–108
22. Cattelan M, Sartori N (2016) Empirical and simulated adjustments of composite likelihood ratio statistics. J Stat Comput Simul 86(5):1056–1067. https://doi.org/10.1080/00949655.2015.1053091
23. Kenkel B (2015) Vectorized bivariate normal cdf. CRAN
24. Bates D, Maechler M (2018) Matrix: sparse and dense matrix classes and methods. R package version 1.2-15. https://CRAN.R-project.org/package=Matrix
25. Wang L, Wu Q (2020) Non-negative variance component estimation for the partial EIV model by the expectation maximization algorithm. Geomat Nat Haz Risk 11(1):1278–1298
26. Cheng SH, Higham NJ (1998) A modified Cholesky algorithm based on a symmetric indefinite factorization. SIAM J Matrix Anal Appl 19(4):1097–1110

27. Medical Expenditure Panel Survey (MEPS) (2008) Content last reviewed august 2018. Agency for Healthcare Research and Quality, Rockville, MD
28. Marra G, Radice R (2019) "gjrm". CRAN
29. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang L-Y, Huang W, Liu B, Shen Y, et al (2003) The international hapmap project
30. Choirat C, Honaker J, Imai K, King G, Lau O (2018) Zelig: Everyone's Statistical Software. Version 5.1.6.1. http://zeligproject.org/
31. Imai K, King G, Lau O (2008) Toward a common framework for statistical analysis and development. J Comput Graph Stat 17(4):892–913