Supplementary Materials for A novel phenotype imputation method with copula model

Jianjun Zhang[1], Jane Zizhen Zhao[2], Samantha Gonzales[3],
Xuexia Wang[3], Qiuying Sha[4*]

[1]Department of Mathematics, University of North Texas, 1155 Union Circle, Denton, 76203, TX, United States.
[2]Department of Psychology and Neuroscience, The University of North Carolina at Chapel Hill, 235 E. Cameron Avenue, Chapel Hill, 27599, NC, United States.
[3]Department of Biostatistics, Florida International University, 11200 S.W. 8th Street, Miami, 33199, FL, United States.
[4]Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, 49931, MI, United States.

*Corresponding author(s). E-mail(s): qsha@mtu.edu;
Contributing authors: Zjj1061219688@163.com; Jan3zhao@unc.edu; samagonz@fiu.edu; xuexwang@fiu.edu;

1 Supplementary Methods

We consider a sample with n unrelated individuals. Each individual has K correlated quantitative traits. Let $\boldsymbol{Y}_i = (y_{i1}, \dots, y_{iK})^T$ denote the phenotype vector for the ith individual, where y_{ik} denotes the kth trait value of the ith individual. We divide the sample into two parts. The first part includes $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{n_1}$ with no missing phenotype. The second part includes $\boldsymbol{Y}_{n_1+1}, \dots, \boldsymbol{Y}_n$ with at least one missing phenotype for each individual. Let $\boldsymbol{Y}_i^{(-K)} = (y_{i1}, \dots, y_{i,K-1})^T$ denote the ith individual's phenotype vector without the K^{th} phenotype. Without loss of generality, we assume that $\boldsymbol{Y}_{n_1+1}^{(-K)}, \dots, \boldsymbol{Y}_n^{(-K)}$ have no missing phenotypes and $Y_{n_1+1,K}, \dots, Y_{n,K}$ have missing values.

We propose to use Gaussian copula to model the correlation among these K traits and let $F_k(y_k; \alpha_k)$ and $f_k(y_k; \alpha_k)$ be the cumulative distribution function (cdf) and probability density function (pdf) of y_k. Usually, we assume y_k follows normal distribution with $N(y_k;\theta_k,\sigma_k^2)$, $\alpha_k=(\theta_k,\sigma_k^2)$, for quantitative traits.

Let $\mu_j=F_j(y_j;\alpha_j), j=1,2...,K$, $C_R=\Phi_R(\Phi^{-1}(\mu_1),\ldots,\Phi^{-1}(\mu_K))$ denotes the joint distribution of $(\mu_1,\ldots,\mu_K)$ where Φ^{-1} is the inverse cumulative distribution function of a standard normal distribution and Φ_R is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix R. Thus, C_R is the cdf of $\boldsymbol{Y}=(y_1,\ldots,y_K)^T$, denoted as $H(y_1,\ldots,y_K)$. Specifically, the distribution of $\boldsymbol{Y}$ will degenerate to a multivariate normal distribution when the marginal distributions of $\boldsymbol{Y}$ are normal based on the Gaussian copula model.

Given the joint distribution function of $\boldsymbol{Y}$, the corresponding density function can be obtained by taking derivatives with respect to C_R [1]. When the trait is continuous, the joint density function of $\boldsymbol{Y}$ can be written as:

$$h(y_1,\ldots,y_K)=c_R(\mu_1,\ldots,\mu_K)\prod_{k=1}^{K}f_k(y_k;\alpha_k) \tag{1}$$

where $c_R(\mu_1,\ldots,\mu_K)=|R|^{-\frac{1}{2}}\exp\{\frac{1}{2}\boldsymbol{q}^T(I_K-R^{-1})\boldsymbol{q}\}$ and $\mu_j=F_j(y_j;\alpha_j)$, $\boldsymbol{q}=(q_1,\ldots,q_K)$ is a vector of inverse-normal scores $q_j=\Phi^{-1}(\mu_j)$, and I_K is a K-dimensional identity matrix. Specially, the conditional density of y_K given $y_1,\ldots,y_{K-1}$ can be written as:

$$h(y_K|y_1,\ldots,y_{K-1})=\frac{c_R(\mu_1,\ldots,\mu_K)f_k(y_k;\alpha_k)}{\int c_R(\mu_1,\ldots,\mu_K)dF_k(y_k;\alpha_k)} \tag{2}$$

When the K traits include K_1 discrete and $K_2=K-K_1$ continuous traits, the joint density function can be obtained as follows. Let $\boldsymbol{\mu_2}=(\mu_{K_1+1},\ldots,\mu_K)$ where $\mu_j=F_j(y_j;\alpha_j)$ for $j=K_1+1,\ldots,K$ and $\mu_{j1}=F_j(y_j-;\alpha_j)$ and $\mu_{j2}=F_j(y_j;\alpha_j)$ where $F_j(y_j-;\alpha_j)$ is the left-hand limit of F_j at y_j which is equal to $F_j(y_j-1;\alpha_j)$ for $j=1,\ldots,K_1$. The joint density of $\boldsymbol{Y}$ is given by:

$$\begin{aligned}h(y_1,\ldots,y_K)=&\prod_{k=K_1+1}^{K}f_k(y_k;\alpha_k)\times\sum_{j_1=1}^{2}\cdots\sum_{j_{K_1}=1}^{2}(-1)^{j_1+\cdots+j_{K_1}}\\&\times C_R^*(\mu_{1,j_1},\cdots,\mu_{K_1,j_{K_1}},\mu_{K_1+1},\ldots,\mu_K)\end{aligned} \tag{3}$$

where

$$\begin{aligned}C_R^*(\boldsymbol{\mu_1},\boldsymbol{\mu_2})=&(2\pi)^{-\frac{K_1}{2}}|R|^{-\frac{1}{2}}\int_{-\infty}^{\Phi^{-1}(\mu_1)}\cdots\int_{-\infty}^{\Phi^{-1}(\mu_{K_1})}\\&\times\exp\left\{-\frac{1}{2}(\boldsymbol{y_1},\boldsymbol{q_2})R^{-1}(\boldsymbol{y_1},\boldsymbol{q_2})^T+\frac{1}{2}\boldsymbol{q_2}^T\boldsymbol{q_2}\right\}d\boldsymbol{y_1}\end{aligned}$$

for $\boldsymbol{\mu_1}=(\mu_1,\ldots,\mu_{K_1})$ and $\boldsymbol{q_2}=(q_{K_1+1},\cdots,q_K)$.

For example, when $K = 3$ and $K_1 = 1$, the joint distribution $\boldsymbol{Y} = (y_1, y_2, y_3)$ can be written as:

$$
\begin{aligned}
&h(y_1, y_2, y_3) = \\
&\prod_{k=2}^{3} f_k(y_k; \alpha_k) \left\{ \int_{-\infty}^{\Phi^{-1}(1-\xi_1)} \frac{1}{\sqrt{2\pi|R|}} \exp\left\{ -\frac{1}{2}(y_1, q_2, q_3) R^{-1} (y_1, q_2, q_3)^T + \frac{1}{2}(q_2, q_3)(q_2, q_3)^T \right\} dy_1 \right\}^{y_1=0} \\
&\times \left\{ 1 - \int_{-\infty}^{\Phi^{-1}(1-\xi_1)} \frac{1}{\sqrt{2\pi|R|}} \exp\left\{ -\frac{1}{2}(y_1, q_2, q_3) R^{-1} (y_1, q_2, q_3)^T + \frac{1}{2}(q_2, q_3)(q_2, q_3)^T \right\} dy_1 \right\}^{y_1=1}
\end{aligned}
\tag{4}
$$

where $\xi_1 = P(y_1 = 1)$ and $1 - \xi_1 = P(y_1 = 0)$. From (4), we can easily derive the conditional probability of y_1 given y_2, y_3:

$$
P(y_1 = 0|y_2, y_3) = \int_{-\infty}^{\Phi^{-1}(1-\xi_1)} \frac{1}{\sqrt{2\pi|R|}} \exp\left\{ -\frac{1}{2}(y_1, q_2, q_3) R^{-1} (y_1, q_2, q_3)^T + \frac{1}{2}(q_2, q_3)(q_2, q_3)^T \right\} dy_1
\tag{5}
$$

and

$$
\begin{aligned}
&P(y_1 = 1|y_2, y_3) = \\
&1 - \int_{-\infty}^{\Phi^{-1}(1-\xi_1)} \frac{1}{\sqrt{2\pi|R|}} \exp\left\{ -\frac{1}{2}(y_1, q_2, q_3) R^{-1} (y_1, q_2, q_3)^T + \frac{1}{2}(q_2, q_3)(q_2, q_3)^T \right\} dy_1
\end{aligned}
\tag{6}
$$

The conditional distribution of y_{iK} given $y_{i1}, \ldots, y_{i(K-1)}$ can be written as:

$$
\begin{aligned}
&h(y_{iK}|y_{i1}, \ldots, y_{i(K-1)}) = \\
&\frac{\sum_{j_1=1}^{2} \cdots \sum_{j_{K_1}=1}^{2} (-1)^{j_1+\cdots+j_{K_1}} C_R^*(\mu_{1,j_1}, \cdots, \mu_{K_1,j_{K_1}}, \mu_{K_1+1}, \ldots, \mu_K) f_K(y_{iK}; \alpha_K)}{\sum_{j_1=1}^{2} \cdots \sum_{j_{K_1}=1}^{2} (-1)^{j_1+\cdots+j_{K_1}} \int_0^1 C_R^*(\mu_{1,j_1}, \cdots, \mu_{K_1,j_{K_1}}, \mu_{K_1+1}, \ldots, \mu_K) d\mu_K}
\end{aligned}
\tag{7}
$$

2 Supplementary Results

We investigate the performance of our methods when the number of phenotypes varies in a broader range from $K = 4$ to $K = 15$ (Supplementary Figure 1). Our methods have similar or better performance than PIM and PHENIX for most of K. Our method performs better than or similarly to PIM and PHENIX when the number of phenotypes ranges from 5 to 14. PHENIX and PIM only outperform our method when the number of phenotypes is 4 or 15. The phenotypes in Supplementary Figure 1 are generated from multivariate normal distributions, which satisfy the key assumptions of PHENIX and PIM. However, when phenotypes are generated from multivariate gamma distributions or a mixture of multivariate normal, multivariate gamma, and beta distributions, the performance of PHENIX and PIM declines even further, as these scenarios violate the key multivariate normal assumption required for their effectiveness (data not shown).

To compare the performance of our methods with the Gaussian copula method using the EM algorithm (Copula-EM) for phenotype imputation [2], three and seven phenotypes are generated from multivariate normal distributions and multivariate gamma distributions, respectively, with $h^2 = 0.05$ and $\rho = 0.5$, while varying the sample size from 100 or 200 to 600.

When the phenotypes follows multivariate normal distributions, our methods outperform Copula-EM consistently for the three-phenotype case and show even greater improvement when the sample size is 100 for the seven-phenotype case. When phenotypes follows multivariate gamma distributions, our methods outperform Copula-EM at larger sample sizes, such as 500 for seven phenotypes or 600 for three phenotypes. This indicates that our methods are particularly robust when dealing with multivariate normal distributions, excelling in both small and large sample sizes. Although our methods are effective for non-normal data, they require a sufficiently large sample size to maintain their advantage over Copula-EM in these situations (Supplementary Figure 2.

References

[1] Li M, Boehnke M, Abecasis G R, and Song PXK. Quantitative trait linkage analysis using Gaussian copulas. Genetics. 2006;173(4):2317–2327.

[2] Zhao, Y. and Udell, M. Missing value imputation for mixed data via Gaussian copula. KDD '20, August 23–27, 2020, Virtual Event. 2020, page 636-646.

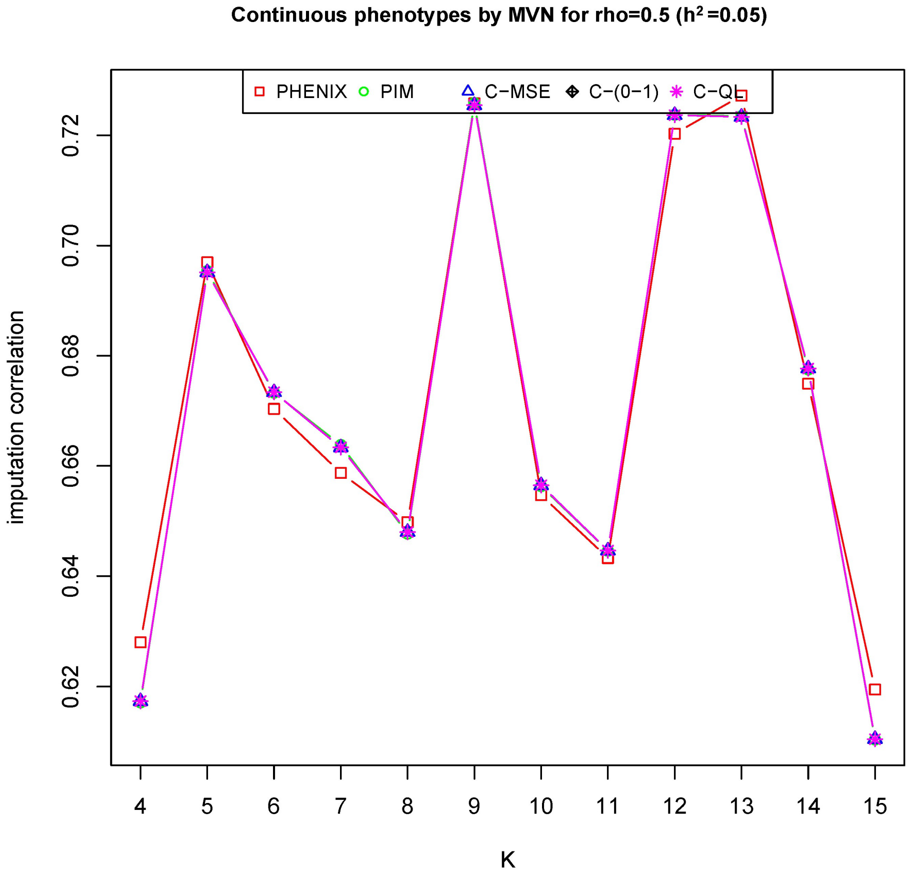
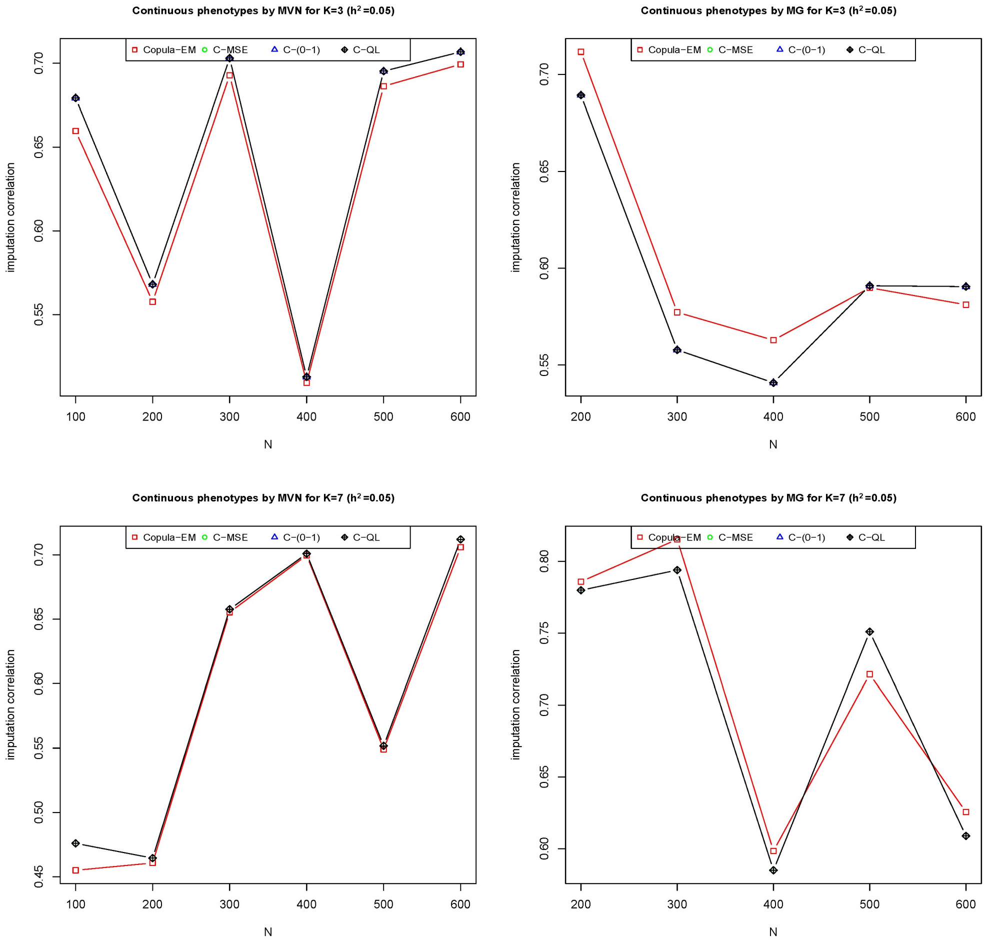


Fig. 1 The imputation correlation of five methods (PHENIX, PIM, C-MSE, C-(0-1) and C-QL) for phenotypes simulated from a multivariate normal distribution at $h^2 = 0.05$, $\rho = 0.5$, $n = 1000$ with a varying number of phenotypes from $K = 4$ to $K = 15$ is shown. The y-axis represents the correlation between the imputed and the true values of phenotypesfor each method.



Continuous phenotypes by MVN for K=3 (h^2=0.05)

Continuous phenotypes by MG for K=3 (h^2=0.05)

Continuous phenotypes by MVN for K=7 (h^2=0.05)

Continuous phenotypes by MG for K=7 (h^2=0.05)

Fig. 2 The imputation correlation of four methods (Copula-EM, C-MSE, C-(0-1) and C-QL) simulated from a multivariate gamma distribution at $h^2 = 0.05$ with different correlation of the traits for a range of positive and negative values where the correlation of the imputed values with the true values is plotted on the y axis for each method.