Principal variables analysis for non-Gaussian data

Dylan Clark-Boucher
dclarkboucher@fas.harvard.edu

Jeffrey W. Miller
jwmiller@hsph.harvard.edu

Abstract

Principal variables analysis (PVA) is a technique for selecting a subset of variables that capture as much of the information in a dataset as possible. Existing approaches for PVA are based on the Pearson correlation matrix, which is not well-suited to describing the relationships between non-Gaussian variables. We propose a generalized approach to PVA enabling the use of different types of correlation, and we explore using Spearman, Gaussian copula, and polychoric correlations as alternatives to Pearson correlation when performing PVA. We compare performance in simulation studies varying the form of the true multivariate distribution over a wide range of possibilities. Our results show that on continuous non-Gaussian data, using generalized PVA with Gaussian copula or Spearman correlations provides a major improvement in performance compared to Pearson. Meanwhile, on ordinal data, generalized PVA with polychoric correlations outperforms the rest by a wide margin. We apply generalized PVA to a dataset of 102 clinical variables measured on individuals with X-linked dystonia parkinsonism (XDP), a rare neurodegenerative disorder, and we find that using different types of correlation yields substantively different sets of principal variables.

1 Introduction

Principal variables analysis (PVA) is technique for selecting a subset of variables that retain the properties of the complete data as well as possible. Different variants of PVA have be used to reduce dimensionality in a way that retains the structure of the original feature space, preserves the relative distance between data points, or explains the variability of features that are not chosen. While principal variables analysis is closely related to principal components analysis (PCA), the main difference is that PVA selects a subset of variables, rather than linear combinations of variables (McCabe, 1984; Beale et al., 1967). In comparison to PCA, PVA is particularly useful for deciding which variables are worth collecting in future studies.

Approaches to PVA have tended to take on three forms. Criterion-based approaches, such as those presented by McCabe (1984) and Cadima et al. (2004), apply best subset algorithms or sequential greedy algorithms to maximize a pre-defined optimality criterion (Brusco, 2014; Cumming and Wooff, 2007). PCA-based approaches, such as those discussed by Jolliffe (1972, 1973), select variables sequentially based on their contributions to important principal components. Clustering-based approaches identify clusters of correlated variables, then select one or more variables from each cluster so that each cluster is represented in the reduced data. Despite their differences, all of these methods share a key feature, which is

that they are based on the sample Pearson correlation matrix. For instance, an appealing criterion-based approach is to minimize the trace of the conditional covariance matrix of the omitted variables given the chosen variables, while holding the number of chosen variables constant (McCabe, 1984). If the variables are multivariate Gaussian, that is, multivariate normal (MVN), and have been normalized to unit variance, this is equivalent to maximizing the "variance explained" — that is, the amount of variance of the omitted variables explained by the chosen variables.

However, a major limitation of these existing methods is that on non-Gaussian data, the Pearson correlation is not a very natural measure of interdependence. For example, it is well-known that two binary variables cannot have correlation 1 unless their marginal probabilities are equal. Similarly, variables may be highly dependent in an unobserved latent space, but exhibit lower correlation in the observed data due to discretization or other transformations that occur in the measurement process. This can cause PVA methods based on the Pearson correlation to miss out on important features.

In this article, we propose a generalized PVA method for handling non-Gaussian data, including ordinal and continuous cases. The basic idea is simply to replace the Pearson correlation matrix with an alternative correlation matrix when running PVA. We investigate using three alternatives: (1) Spearman correlation, which is based on the observed ranks; (2) polychoric correlation, which is designed for ordinal variables (Choi et al., 2010); and (3) Gaussian copula correlation, which can capture arbitrary latent correlation structure and arbitrary marginals (Trivedi and Zimmer, 2005). We evaluate each approach under diverse simulation conditions with varying assumptions about the true multivariate distribution. We focus on the criterion-based method for PVA described above, in which we seek to minimize the trace of the conditional covariance matrix of the omitted variables given the chosen variables. To the best of our knowledge, this is the first study to directly examine the limitations of using the Pearson correlation matrix for PVA, while presenting alternatives that are robust to changes in the true data-generating mechanism.

The article is organized as follows. In Section 2, we describe the PVA algorithm based on the residual trace criterion, and we extend the method beyond Pearson to Spearman, polychoric, and Gaussian copula correlations. In Section 3, we evaluate the performance of these methods on Gaussian and non-Gaussian data in a range of simulation studies. In Section 4, we apply the methods to a dataset containing 102 clinical variables from individuals with X-linked dystonia parkinsonism (XDP), a rare and under-studied genetic disease. We conclude in Section 5 with a brief discussion.

2 Methods

2.1 PVA algorithm

The variant of PVA we will consider is based on the McCabe "variance-explained" criterion, which selects the subset of variables that explains as much of the variance in the other variables as possible. Let X denote a random vector of length p and let $S \subseteq \{1, \ldots, p\}$ such that $|S| = q$. We write X_S to denote the sub-vector $X_S = (X_j : j \in S) \in \mathbb{R}^q$, and X_{S^c} to denote the complementary sub-vector $X_{S^c} = (X_j : j \notin S) \in \mathbb{R}^{p-q}$. Then the optimal subset

is defined as

$$S^* = \underset{S\,:\,|S|=q}{\operatorname{argmin}} \operatorname{tr}\big(\operatorname{Cov}(X_{S^c} \mid X_S)\big) \tag{1}$$

where $\operatorname{tr}(\cdot)$ is the trace of a matrix. Other criteria for PVA have considered minimizing the determinant of $\operatorname{Cov}(X_{S^c} \mid X_S)$ or the Frobenius norm (McCabe, 1984; Cumming and Wooff, 2007).

The covariance matrix in Equation (1) is straightforward to compute when the data follow a multivariate Gaussian distribution. If $[X_S^\intercal, X_{S^c}^\intercal]^\intercal \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

such that $\Sigma_{11} = \operatorname{Cov}(X_S)$ and $\Sigma_{22} = \operatorname{Cov}(X_{S^c})$, then the conditional covariance matrix of X_{S^c} given X_S is

$$\boldsymbol{\Sigma}_{S^c|S} := \operatorname{Cov}(X_{S^c} \mid X_S) = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \tag{2}$$

As a special case of this formula, if X_j is the jth entry of X, and $-j$ denotes $\{j\}^c$, then the conditional covariance of all the other variables in X given X_j is given by

$$\boldsymbol{\Sigma}_{-j|j} := \operatorname{Cov}(X_{-j} \mid X_j) = \boldsymbol{\Sigma}_{-j} - \boldsymbol{\Sigma}_{-j,j}\boldsymbol{\Sigma}_{-j,j}^\intercal/\sigma_j^2, \tag{3}$$

where $\sigma_j^2 = \operatorname{Var}(X_j)$, $\boldsymbol{\Sigma}_{-j} = \operatorname{Cov}(X_{-j})$, and $\boldsymbol{\Sigma}_{-j,j}$ is a vector of the covariance between each entry of X_{-j} and X_j.

To avoid a computationally intensive search over sub-vectors of X, we employ a greedy algorithm for finding an approximate solution to the optimization problem; see Algorithm 2.1. Briefly, starting with an estimated covariance matrix, the first step in the algorithm is to find the variable X_j that yields a conditional covariance matrix with the smallest trace based on Equation (3). This j becomes the first index j_1 to be included in our selected set $\tilde{S}^* = \{j_1, \ldots, j_q\}$, and the algorithm repeats with $\boldsymbol{\Sigma}_{-j|j}$ as the new covariance matrix.

Algorithm 2.1.

Input: $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ *positive definite.*

Output: Indices of selected variables, $j_1, \ldots, j_q$.

(1) Initialize $S \leftarrow \{1, \ldots, p\}$.

(2) For $k = 1, \ldots, q$*:*

(a) $i \leftarrow \operatorname{argmin}_j \operatorname{tr}(\boldsymbol{\Sigma}_{-j|j})$

(b) $j_k \leftarrow S_i$

(c) $\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma}_{-i|i}$

(d) $S \leftarrow S \setminus \{j_k\}$.

Note that this algorithm can be run without directly observing the data, since the only required input is a matrix $\boldsymbol{\Sigma}$. In addition, since the scale of each variable is not relevant to the dependency between variables, a natural input to this algorithm is the correlation matrix rather than the covariance matrix. In the next section, we discuss the limitations of using the Pearson correlation matrix for this purpose and consider more robust alternatives for the choice of $\boldsymbol{\Sigma}$.

2.2 Generalized PVA using alternative correlations

It is widely recognized that Pearson correlation has drawbacks when measuring the interdependence of non-Gaussian data. For example, it is not invariant to monotone transformations of the data, it is constrained by the marginal distributions, and it is incapable of describing bivariate tail dependence. We consider three alternatives to Pearson correlation in the context of PVA: Spearman correlation, polychoric correlation, and Gaussian copula correlation.

Spearman correlation is equivalent to the Pearson correlation of the ranks. Ties are handled by averaging the ranks of each set of tied points and assigning this average rank to those points, which is equivalent to averaging over all permutations of the ties (Dodge, 2008). An attractive feature of Spearman correlation is that it is invariant to monotonically increasing transformations of the data; thus, it is nonparametric with respect to the marginal distributions. However, since the ranks of the data are uniformly distributed, the conditional covariance is not given by the formula in Equation (2), since this formula is based on the multivariate Gaussian assumption.

Polychoric correlation extends the Pearson correlation to ordinal variables by assuming the variables have a latent multivariate Gaussian structure (Choi et al., 2010). Specifically, suppose $[Z_1, Z_2]^\intercal \in \mathbb{R}^2$ is bivariate Gaussian with unit variances and correlation ρ, and Y_1, Y_2 are ordinal variables defined by unknown monotonic transformations of Z_1, Z_2, respectively. Then ρ is referred to as the polychoric correlation of Y_1, Y_2, and it can be estimated via an iterative maximum-likelihood approach given observations of Y_1, Y_2 (Olsson, 1979). The estimation procedure can also be applied when we have observations of Z_1, Y_2 rather than Y_1, Y_2.

Gaussian copulas are tractable models of multivariate dependence with arbitrary marginal distributions, popular in economics (Trivedi and Zimmer, 2005). Suppose the data comprise n observations of a multivariate random vector $X = (X_1, X_2, \ldots, X_p)^T$, with an unknown cumulative distribution function $F_X(x_1, x_2, \ldots, x_p)$. A Gaussian copula provides an approximation to F_X of the form

$$\hat{F}_X(x_1, x_2, \ldots, x_p) = \Phi_p\big(\Phi^{-1}(\hat{F}_1(x_1)), \Phi^{-1}(\hat{F}_2(x_2)), \ldots, \Phi^{-1}(\hat{F}_p(x_p)) \;\big|\; \mathbf{0}, \hat{\boldsymbol{\Sigma}}\big),$$

where Φ is the cumulative distribution function (CDF) of a standard normal distribution, $\Phi_p(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the CDF of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_p$ are estimated CDFs of the p variables. To fit this model, the approximations $\hat{F}_j$ can be obtained parametrically by specifying a functional form of the distributions and applying maximum likelihood, or non-parametrically by taking the empirical CDF of the observed data. Finally, the estimate $\hat{\boldsymbol{\Sigma}}$ can be defined as the sample Pearson correlation matrix of the transformed variables $\Phi^{-1}(\hat{F}_1(X_1)), \Phi^{-1}(\hat{F}_2(X_2)), \ldots, \Phi^{-1}(\hat{F}_p(X_p))$. The interpretation of $\hat{\boldsymbol{\Sigma}}$ is the covariance matrix of a multivariate Gaussian distribution that can be constructed by applying monotonic transformations to the observed variables. However, the validity of this interpretation depends on the strong assumption that the transformed variables $\Phi^{-1}(F_1(X_1)), \Phi^{-1}(F_2(X_2)), \ldots, \Phi^{-1}(F_p(X_p))$ are multivariate Gaussian.

Each of these three approaches (Spearman, polychoric, and Gaussian copula) produces an estimated correlation matrix $\boldsymbol{\Sigma}$ that can be used to perform PVA using Algorithm 2.1. Importantly, the conditional covariance formula in Equation (2) is valid for the polychoric

and Gaussian copula approaches, since $\mathbf{\Sigma}$ is the correlation matrix of multivariate Gaussian latent variables in these approaches. The implied assumption of these approaches is that we are interested in the dependence structure of the latent variables rather than the observed variables, which may or may not be appropriate depending on the application. In the application in Section 4, for instance, this is a natural assumption since the ordinal variables represent continuous traits that were measured on ordinal scales for practical reasons.

3 Simulations

In this section, we evaluate the performance of PVA using the Spearman, Gaussian copula, and polychoric correlation approaches compared to the standard approach using the Pearson correlation. To simulate data, we generate latent variables from a multivariate Gaussian distribution and transform them by applying monotonic functions to each variable separately, in order to evaluate the effect of non-Gaussianity on PVA performance. The latent multivariate Gaussian distribution serves as ground truth for defining the ideal value of the optimality criterion that would be used if the latent distribution were known. We consider a variety of monotonic transformations (yielding a variety of non-Gaussian marginal distributions) and we assess performance in two ways: (1) the proportion of ideal variables selected, and (2) the variance explained by the selected variables, quantified in terms of relative explanatory efficiency.

3.1 Proportion of ideal variables selected

For the first set of simulations, we consider examples in which non-Gaussianity causes Pearson correlation-based PVA to miss variables that are important for capturing latent multivariate structure. The general setup is to simulate a latent data matrix $\boldsymbol{X} = [X_{ij}] \in \mathbb{R}^{n \times p}$ comprising n points sampled i.i.d. from a $\text{MVN}_p(\mathbf{0}, \mathbf{\Sigma})$ distribution, and determine an "ideal" set of q latently important variables by performing PVA on $\mathbf{\Sigma}$. We then generate $\boldsymbol{Y}$, the observed data matrix, by the transformation

$$Y_{ij} = \begin{cases} f_j(X_{ij}) & \text{if } j \in H \\ X_{ij} & \text{if } j \notin H \end{cases}$$

where $f_1, \ldots, f_p$ are monotonic functions and $H \subseteq \{1, 2, \ldots, p\}$ is a fixed subset of K variables that will be transformed. We evaluate the performance of PVA with different correlation matrices (Pearson, Spearman, Gaussian copula, and polychoric) estimated from the data $\boldsymbol{Y}$ by contrasting the sets of variables obtained from those methods with the ideal set of variables obtained by PVA on $\mathbf{\Sigma}$. The more these two sets of variables overlap, the better PVA is performing.

We run simulations using the following settings. With $p = 10$, we randomly generate $\mathbf{\Sigma}$ by sampling from a Wishart distribution with p degrees of freedom and scale matrix I, then normalize this matrix to have ones on the diagonal so it is a correlation matrix. We then perform PVA on this ground truth $\mathbf{\Sigma}$ to identify the ideal set of $q = 5$ variables, say $j_1^*, \ldots, j_5^*$, that capture the most variance in the latent multivariate distribution. We then

sample latent data $\boldsymbol{X}$ from $\text{MVN}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, and generate the observed data $\boldsymbol{Y}$ by transforming variables $H = \{j_1^*, \ldots, j_5^*\}$ in one of three different ways: (A) no transformation, (B) continuous transformation, or (C) ordinal transformation. (That is, the five variables that are most important latently are the variables that are transformed). In each case, for each $n \in \{50, 150, 400, 1200, 3500, 10000\}$, we perform 1000 replicate simulations of the process of generating $\boldsymbol{\Sigma}, \boldsymbol{X}, \boldsymbol{Y}$, running the PVA methods, and comparing the chosen sets of variables to the latently ideal set, $\{j_1^*, \ldots, j_5^*\}$.

(A) *No transformation.* To compare performance when the data are indeed multivariate Gaussian, we run the suite of simulations without any transformation of $\boldsymbol{X}$, so that $\boldsymbol{Y} = \boldsymbol{X}$. We run PVA using Pearson, Spearman, and Gaussian copula correlations; polychoric is not considered here since it is designed for ordinal variables.

(B) *Continuous transformation.* In this case, we transform the ideal variables $j_1^*, \ldots, j_5^*$ with the five monotonically increasing functions

$$\begin{aligned}
Y_{ij_1^*} &= \min(\hat{F}_{j_1^*}(X_{ij_1^*})^2, 0.6^2) \\
Y_{ij_2^*} &= \min(\hat{F}_{j_2^*}(X_{ij_2^*})^6, 0.6^6) \\
Y_{ij_3^*} &= F^{-1}_{\text{Gamma}}(\hat{F}_{j_3^*}(X_{ij_3^*}), 0.5, 1) \\
Y_{ij_4^*} &= F^{-1}_{\text{Pareto}}(\hat{F}_{j_4^*}(X_{ij_4^*}), 2, 1) \\
Y_{ij_5^*} &= \exp\left(\hat{F}_{j_5^*}(X_{ij_5^*})\right)[1 + \mathbb{1}(\hat{F}_{j_5^*}(X_{ij_5^*}) > 0.9)]
\end{aligned}$$

for $i \in \{1, 2, \ldots, n\}$, where $\hat{F}_k$ is the empirical CDF of $X_{1k}, \ldots, X_{nk}$, times $n/(n+1)$, so that

$$\hat{F}_k(X_{ik}) = \frac{\text{Rank}(X_{ik})}{n+1}.$$

Here, $\text{Rank}(X_{ik})$ is the ranking (from least to greatest) of X_{ik} in $\boldsymbol{X}_k$, the kth column of $\boldsymbol{X}$. (Ties in the rank are handled by taking the average rank over all permutations of the column.) Meanwhile, $\mathbb{1}(\cdot)$ is the indicator function, $F^{-1}_{\text{Gamma}}(\cdot, a, b)$ is the inverse CDF of the gamma distribution with shape a and scale b, and $F^{-1}_{\text{Pareto}}(\cdot, a, b)$ is the inverse CDF of a Pareto distribution with shape a and scale b. The empirical CDFs transform the variables to be approximately Uniform$(0, 1)$, marginally, after which they are mapped through an inverse CDF to produce one of five marginal distributions. These inverse CDFs are chosen to reduce the Pearson correlation between the ideal variables and the non-ideal variables, while creating marginal distributions that could realistically be observed in practice. For these simulations, we again use Pearson, Spearman, and Gaussian copula correlations, but not polychoric since it is designed for ordinal variables.

(C) *Ordinal transformation.* In this case, we transform the ideal variables $j_1^*, \ldots, j_5^*$ from Gaussian to ordinal variables by applying the following functions, respectively:

$$\begin{aligned}
Y_{ij_1^*} &= 1 + \mathbb{1}(\hat{F}_{j_1^*}(X_{ij_1^*}) > 0.2) \\
Y_{ij_2^*} &= 1 + \mathbb{1}(\hat{F}_{j_2^*}(X_{ij_2^*}) > 0.4) + \mathbb{1}(\hat{F}_{j_2^*}(X_{ij_2^*}) > 0.6) \\
Y_{ij_3^*} &= 1 + \mathbb{1}(\hat{F}_{j_3^*}(X_{ij_3^*}) > 0.2) + \mathbb{1}(\hat{F}_{j_3^*}(X_{ij_3^*}) > 0.3) \\
Y_{ij_4^*} &= 1 + \mathbb{1}(\hat{F}_{j_4^*}(X_{ij_4^*}) > 0.3) + \mathbb{1}(\hat{F}_{j_4^*}(X_{ij_4^*}) > 0.5) + \mathbb{1}(\hat{F}_{j_4^*}(X_{ij_4^*}) > 0.7) \\
Y_{ij_5^*} &= 1 + \mathbb{1}(\hat{F}_{j_5^*}(X_{ij_5^*}) > 0.1) + \mathbb{1}(\hat{F}_{j_5^*}(X_{ij_5^*}) > 0.2) + \mathbb{1}(\hat{F}_{j_6^*}(X_{ij_5^*}) > 0.3).
\end{aligned}$$

This makes $Y_{ij_1^*} \in \{1, 2\}$, $Y_{ij_2^*}, Y_{ij_3^*} \in \{1, 2, 3\}$, and $Y_{ij_4^*}, Y_{ij_5^*} \in \{1, 2, 3, 4, 5\}$. Here, we run PVA with polychoric correlations in addition to Pearson, Spearman, and Gaussian copula, since polychoric correlations are designed specifically for the case of ordinal variables.

Figure 1: Proportion of latently ideal variables identified by PVA. Value shown is the mean proportion over 1,000 replicate simulations, and error bars show ± two standard errors. The x-axis is on a logarithmic scale.

3.1.1 Results for proportion of ideal variables selected

Results for these simulations are shown in Figure 1. The y-axis shows the mean proportion of the latently ideal variables that were selected by PVA when using each method (Pearson, Spearman, Gaussian copula, or polychoric). In case A (no transformation), where the observed data are Gaussian, there is effectively no difference between the performance of the Pearson and Gaussian copula correlation approaches; both perform very well with the proportion nearing 100% as the sample size exceeds 10,000. Spearman correlations also perform well but slightly worse. This makes sense since Pearson and Gaussian copula correlations either implicitly or explicitly employ Gaussian assumptions, whereas Spearman is a fully nonparametric approach that is expected to entail some loss of information.

In case B (continuous transformation), Spearman and Gaussian copula correlations perform moderately well—with the proportion of ideal variables chosen exceeding 80%—whereas Pearson is unable to reach even 60%. This accords with intuition since the Spearman and

Gaussian copula approaches are designed to handle non-Gaussian marginals, whereas Pearson is not robust to departures from Gaussianity. The performance of Pearson correlations improves slightly with increasing n, but appears to plateau below 60%.

Finally, in case C (ordinal transformation), polychoric correlations are by far the best option, roughly similar in efficacy to the best performing methods in the "no transformation" case. Meanwhile, Pearson, Spearman, and the Gaussian copula struggle in the ordinal case. These results make sense because polychoric correlation is designed for ordinal variables, whereas Pearson is best suited for Gaussians, and the Spearman and Gaussian copula approaches likely break down due to the large number of ties.

Among these simulations, the only case in which PVA with Pearson correlation performs adequately is when the data are truly multivariate Gaussian, where it performs similarly to Gaussian copulas and slightly better than Spearman correlations. This indicates that the standard PVA approaches in the literature are not well suited to non-Gaussian data when we are interested in the latent relationships among variables, and that significantly better performance can be obtained with the alternative methods we propose.

3.2 Relative explanatory efficiency

A limitation of measuring performance in terms of the proportion of ideal variables selected is that the amount of overlap between a given set of variables and the ideal set of variables is not necessarily indicative of their relative utility, since the correlations among variables may be complex. For example, it could happen that replacing two of the ideal variables with two non-ideal variables would yield a set with greater utility than replacing either one of the two variables alone, even though this would yield a set with a smaller proportion of the ideal variables.

Thus, we also consider the utility of chosen sets of variables in terms of how much of the variance of the omitted variables they explain. Borrowing from the language of "relative efficiency" of statistical estimators, we introduce the "relative explanatory efficiency" of X_S versus X_{S^*}, defined as

$$\mathrm{REE}(X_S, X_{S^*}) = \frac{\mathrm{tr}(\mathrm{Cov}(X \mid X_{S^*}))}{\mathrm{tr}(\mathrm{Cov}(X \mid X_S))}, \tag{4}$$

where X_S and X_{S^*} are potentially overlapping sub-vectors of the random vector X. We can interpret $\mathrm{tr}(\mathrm{Cov}(X \mid X_S))$ as the total remaining marginal variance of the variables in X conditional on X_S; note that $\mathrm{tr}(\mathrm{Cov}(X \mid X_S)) = \mathrm{tr}(\mathrm{Cov}(X_{S^c} \mid X_S))$ since the variables in X_S have conditional variance 0. Thus, if $\mathrm{REE}(X_S, X_{S^*}) > 1$ then this indicates that X_S is superior to X_{S^*} in terms of variance explained across all variables, whereas $0 < \mathrm{REE}(X_S, X_{S^*}) < 1$ indicates the reverse.

Our second suite of simulations evaluates the REE of the chosen sets of variables X_S versus the ideal set X_{S^*}. We repeat our data-generating set up from Section 3.1, but use REE as the performance metric instead of the percent the ideal variables chosen. Like before, we perform 1000 replicated simulations in which $\mathbf{\Sigma}$ is sampled from a Wishart distribution, with $p = 10$ degrees of freedom and scale matrix $\boldsymbol{I}$. Once $\mathbf{\Sigma}$ has been converted to a correlation matrix, we produce the latent data $\boldsymbol{X} = [X_{ij}] \in \mathbb{R}^{n \times p}$ by sampling n observations from a $\mathrm{MVN}_p(\mathbf{0}, \mathbf{\Sigma})$ distribution. The observed data $\boldsymbol{Y} = [Y_{ij}] \in \mathbb{R}^{n \times p}$ are obtained by transforming

the 5 ideal latent variables, X_{S^*}, by the transformations A, B, or C from Section 3.1, where X_{S^*} is determined by performing PVA on $\mathbf{\Sigma}$. Finally, the set of "chosen" variables are determined by applying PVA to $\boldsymbol{Y}$ using either the Pearson, Spearman, Gaussian copula, or polychoric correlation matrix. (Polychoric is applied only for transformation C). We compare the chosen set of variables to the ideal set of variables using Equation (4). The procedure is repeated for $n \in \{50, 150, 400, 1200, 3500, 10000\}$.

3.2.1 Results for relative explanatory efficiency

Results for these simulations are shown in Figure 2. The y-axis shows the mean REE of the chosen variables relative to the ideal variables, expressed as a percentage. In case A (no transformation), the Pearson correlation method performs slightly better than the Spearman and Gaussian copula method when is n is small. But as n grows large, the mean REE of each of the methods appears to converge to 100%, and the gaps between the methods disappear. In case B (continuous transformation), the mean REE for the Pearson correlation method plateaus between 95% and 97.5%, while the mean REE for the Spearman and Gaussian copula methods again approaches 100%. In case C (ordinal transformation), the Pearson, Spearman, and Gaussian copula approaches struggle for all choices of n, with their mean REEs plateauing below 97.5%. In contrast, the polychoric method, which is designed specifically for ordinal variables, yields close to 100% REE with large enough sample size, similar to how the other methods performed when there was no transformation. In general, our findings here are similar to those observed in Figure 1, where we evaluated performance in terms of the proportion of ideal variables selected rather than REE. However, the performance gaps between methods appear smaller when using REE, since non-ideal sets of variables may still be effective at capturing latent variance.

3.3 Expanded suite of simulations

Next, we consider an expanded suite of simulations for generating and analyzing the data. Specifically, we perform simulations under an array of settings in which (1) the number of variables chosen, q, can vary; (2) the latent distribution that underlies the data is not necessarily Gaussian, and (3) every latent variable is transformed rather than just the ideal variables.

In this expanded suite of simulations, we again sample a latent correlation matrix $\mathbf{\Sigma}$ that will be used to define an "ideal" set of variables for that specific q. In addition to multivariate Gaussian latent variables, we also consider cases in which the latent variables follow either a multivariate t (MVT) or a multivariate generalized Laplace (MVL) distribution. The MVT and MVL distributions are convenient for this purpose because they are mixtures of the multivariate Gaussian distribution with certain univariate distributions, meaning they have a correlation matrix as one of their parameters. Specifically, if $W \in \mathbb{R}^p \sim \text{MVN}_p(\mathbf{0}, \mathbf{\Sigma})$ and $Z \in \mathbb{R} \sim \mathcal{X}^2_\nu$ independently, then $W/\sqrt{Z/\nu} \sim \text{MVT}_p(\nu, \mathbf{0}, \mathbf{\Sigma})$. Meanwhile, if $W \in \mathbb{R}^p \sim \text{MVN}_p(\mathbf{0}, \mathbf{\Sigma})$ and $Z \in \mathbb{R} \sim \text{Gamma}(r, 1)$ independently, then $W\sqrt{Z} \sim \text{MVL}_p(r, \mathbf{\Sigma})$. These distributions differ from the multivariate Gaussian not only in their marginal distributions but also in their dependence structure, since the mixing with Z introduces tail dependence such that an extreme value in one entry of $\boldsymbol{X}$ increases the probability of extreme values

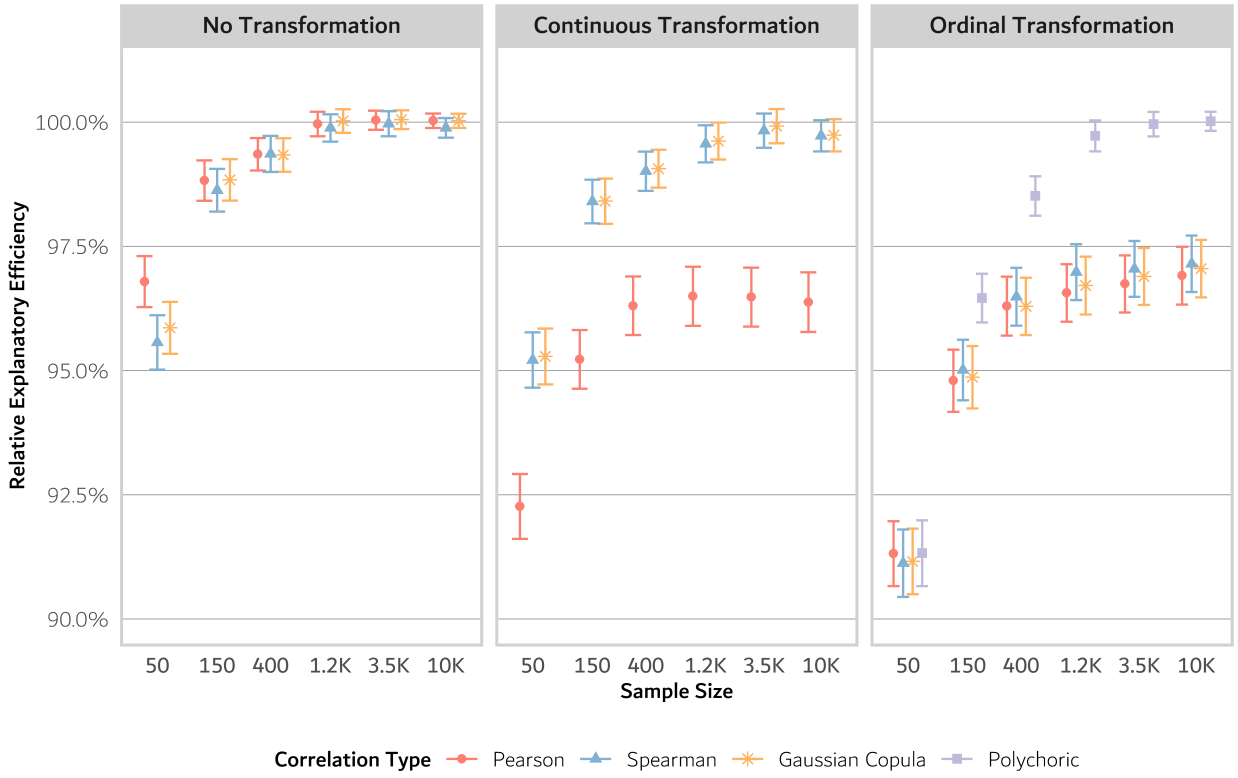

Figure 2: Relative explanatory efficiency of the chosen variables compared to the ideal variables. The y-axis is the variance unexplained by the ideal variables divided by the variance unexplained by the selected variables, as a percentage. The value shown is the average over 1,000 simulations, and the error bars show ± two standard errors. Note that REE $> 100\%$ is mathematically possible, since the greedy algorithm is not guaranteed to find the global optimum.

in the other entries, which is not the case for multivariate Gaussians. Unfortunately, for these distributions, it is more difficult to compute $\mathrm{Cov}(X \mid X_S)$ and, further, $\mathrm{Cov}(X \mid X_S)$ depends on the specific value taken by X_S rather than just depending on S and $\boldsymbol{\Sigma}$. As a workaround, we modify the PVA procedure in Algorithm 2.1 by replacing $\boldsymbol{\Sigma}_{-j|j}$ with $\boldsymbol{\Sigma}^0_{-j|j} := \mathrm{Cov}(X_{-j} \mid X_j = \mathrm{E}X_j) = \mathrm{Cov}(X_{-j} \mid X_j = 0)$, the conditional covariance given that the selected variable takes its mean value. For the $\mathrm{MVT}_p(\nu, \mathbf{0}, \boldsymbol{\Sigma})$ case,

$$\mathrm{Cov}(X_{-j} \mid X_j = 0) = \frac{\nu}{\nu - 1}(\boldsymbol{\Sigma}_{-j} - \boldsymbol{\Sigma}_{-j,j}\boldsymbol{\Sigma}^{\mathsf{T}}_{-j,j}/\Sigma_{jj}) \tag{5}$$

by Dodge (2008), and for the $\mathrm{MVL}_p(r, \boldsymbol{\Sigma})$ case,

$$\mathrm{Cov}(X_{-j} \mid X_j = 0) = (r - 0.5)(\boldsymbol{\Sigma}_{-j} - \boldsymbol{\Sigma}_{-j,j}\boldsymbol{\Sigma}^{\mathsf{T}}_{-j,j}/\Sigma_{jj}) \tag{6}$$

by Kotz et al. (2001) and Kozubowski et al. (2013). These formulae can be adapted to induce a corresponding version of REE, in which $\mathrm{tr}(\mathrm{Cov}(X \mid X_S = \mathbf{0}_q))$ replaces $\mathrm{tr}(\mathrm{Cov}(X \mid X_S))$.

The structure of the expanded simulations is similar to previously. Once the correlation matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ has been sampled ($p = 10$), we generate the latent data $\boldsymbol{X}$ by sampling n

observations independently from either an MVN($\mathbf{0}_p$, $\boldsymbol{\Sigma}$), MVT($\nu = 2.5$, $\boldsymbol{\Sigma}$), or MVL($r = 3.1$, $\boldsymbol{\Sigma}$) distribution. We produce the observed data, $\boldsymbol{Y}$, by transforming each of the p variables by the monotonic transformations (A, B, or C) that were defined in Section 3. Since $p = 10$, each of the five transformations is used twice. PVA is then performed to select the best set of q variables, X_S, using either the Pearson, Spearman, Gaussian copula, or polychoric correlation matrix based on $\boldsymbol{Y}$. We evaluate the performance of each method in terms of REE(X_S, X_{S^*}), where X_{S^*} is determined by PVA on $\boldsymbol{\Sigma}$. The procedure is repeated 1000 times for each $q \in \{2, 3, 4, 5, 6\}$, fixing $n = 500$. The results of this procedure are provided below. In addition, for completeness, we repeat our analyses from Sections 3.1 and 3.2 in the cases of multivariate t or multivariate Laplace latent data; see Appendix A.

3.3.1 Results for expanded suite of simulations

Figure 3 shows the results. Each row of the figure is for a different latent distribution, and each column shows a different transformation type. The y-axis of each plot is the relative explanatory efficiency (REE) of the chosen set of variables compared to the ideal set as determined by running PVA on $\boldsymbol{\Sigma}$. We report the average REE over 1000 simulations.

When the latent distribution is multivariate Gaussian (Figure 3A) and the variables are not transformed (left panel), there are no discernible differences between the three correlation methods (Pearson, Spearman, and Gaussian copula): all three methods perform excellently, with relative efficiencies close to 100%. However, the performance of the Pearson method is negatively affected by nonlinear transformations, as seen in the "continuous transformation" and "ordinal transformation" cases (Figure 3A, middle and right). The Spearman and Gaussian copula methods behave similarly to one another no matter the transformation type, performing well when there are continuous transformations but poorly when the data are ordinal. The ordinal case is best handled by the polychoric method, which clearly outperforms the other methods. Empirically, we observe that REE decreases with the number of variables chosen, indicating that the relative performance of the ideal set compared to the chosen set becomes more prominent as q grows larger.

Figure 3B and Figure 3C show the results when the latent distribution is multivariate t or multivariate Laplace, respectively. For the multivariate t case, the REEs of the chosen sets of variables compared to the ideal sets of variables are consistently lower than the results in the multivariate Gaussian case, across the board, with the Pearson correlation method appearing to suffer the most. The reduced performance of Pearson compared to the other methods occurred even when the data were not transformed, which emphasizes the lack of robustness of Pearson correlations to violations of multivariate normality. In contrast, the Spearman and Gaussian copula methods were more robust to the change in latent distribution, exhibiting only slightly decreased REEs. The polychoric method was again the best performer when the variables were transformed to be ordinal. Finally, in the multivariate Laplace case, the REEs were only slightly lower across the board compared to the multivariate Gaussian case, with all the methods being affected similarly. In particular, the Pearson REEs were only slightly reduced in the multivariate Laplace case.

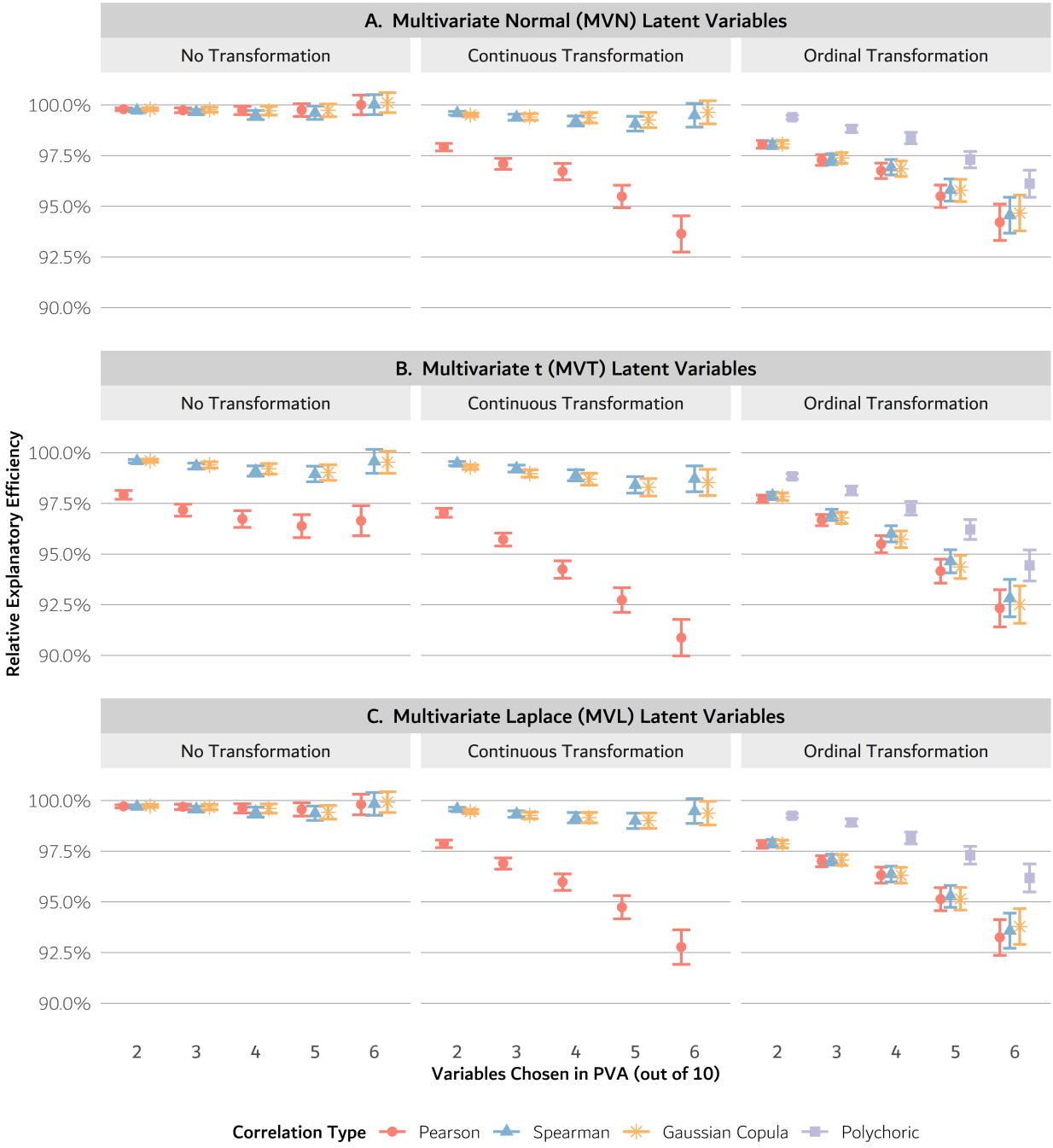

Figure 3: Relative explanatory efficiency in the expanded suite of simulations. The y-axis is the variance unexplained by the ideal variables divided by the variance unexplained by the selected variables, as a percentage. The value shown is the average over 1,000 simulations, and the error bars show ± two standard errors. Note that REE > 100% is mathematically possible.

4 Application

To compare PVA methods on real data, we consider a dataset of longitudinal clinical measurements from individuals with X-linked dystonia parkinsonism (XDP), a rare neurodegenerative disorder occurring in men with maternal ancestry from the island of Panay, Phillipines (Lee et al., 1976). As the name suggests, XDP involves symptoms characteristic of both dystonia (such as muscle contractions, repetitive movements, and abnormal posture) and parkinsonism (such as bodily tremors, slowed movement, and muted facial expressions) (Lee et al., 2011). XDP is strongly associated with the insertion of a SINE-VNTR-Alu (SVA) retrotransposon in the *TAF1* gene (Aneichyk et al., 2018; Makino et al., 2007). The SVA retrotransposon contains repeated nucleotide sequences of the form CCCTCT, and greater numbers of repeats of this sequence (that is, larger "repeat size") are associated with earlier ages of onset (Bragg et al., 2017). In general, XDP symptoms begin in middle age and then gradually worsen over time.

In preparation for future clinical trials, Acuna et al. (2023) conducted a longitudinal study of XDP symptoms, using Pearson correlation-based PVA to help define a minimal battery of clinical measures that could be used to assess the efficacy of a treatment for slowing or reversing XDP symptoms. Acuna et al. (2023) collected data on 29 symptomatic adult males with XDP, visited at 6-month intervals during an 18-month span. Symptoms were assessed on several published clinical scales, including the Movement Disorders Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS), which measures parkinsonism traits; the Burke-Fahn-Marsden Dystonia Rating Scale (BFM), which measures dystonia-related symptoms; the Eating Assessment Tool (EAT-10), which measures swallowing impairment; and the Communicative Participant Item Bank (CPIB), which measures difficulty communicating. Other measurements targeted the tongue strength and lip strength of the patients, and their ability to make and repeat sounds common to their native language. See Acuna et al. (2023) for further details.

Our analysis focuses on the 29 XDP-positive individuals who were studied by Acuna et al. (2023), each of whom had measurements taken at up to five distinct visits, amounting to 105 visits in total. The variables we consider comprise 102 clinical measures taken at each of those visits; 17 of the variables are continuous and 85 of them are ordinal, with the ordinal variables taking anywhere from 3 to 10 unique levels. The visits are treated independently and serve as the experimental units of the analysis (that is, $n = 105$). We handle missingness among the measurements by generating a singly-imputed dataset using predictive mean matching, for which we use the R package MICE (Rubin, 1986; van Buuren and Groothuis-Oudshoorn, 2011). Following the imputation, we run PVA using Pearson, Spearman, Gaussian copula, and polychoric correlations, and compare their performance in terms of finding a small subset of variables that capture as much as possible of the information present in all 102 measures, that is, in terms of defining a minimal battery for future studies.

4.1 Application results

Table 1 shows the first ten variables chosen by performing PVA based on the Pearson, Spearman, Gaussian copula, and polychoric correlation matrices, respectively. We also tried applying polychoric correlations with the continuous variables transformed to be marginally

uniform or marginally normal, but this did not make any difference in the first ten variables chosen.

In the first column of Table 1, we see that PVA based on Pearson correlations chose a variety of variables describing attributes such as speech, swallowing, walking, rigidity, and tremors—covering both dystonia symptoms (the BFM measures) and parkinsonism symptoms (the UPDRS measures). Meanwhile, in the second column, PVA based on Spearman correlations had a strong preference for UPDRS measures over BFM measures; in fact, the only two non-UPDRS measures selected were reported age at onset and EAT-10 Q2. In the third column, we see that the Gaussian copula approach tended to favor UPDRS measures over BFM measures as well, choosing 7 UPDRS measures and only 1 BFM measure. The Gaussian copula was also the only method to select SVA repeat size, which is a known predictor of XDP age at onset. Finally, the fourth column shows results for polychoric correlations, which like the Spearman and Gaussian copula approaches, strongly preferred UPDRS variables instead of BFM variables, although one BFM variable was chosen tenth. Using Gaussian copula or polychoric correlations also resulted in the omission of age at onset, which both the Pearson and Spearman methods did select, albeit close to last. All four methods selected at least one variable describing difficulty eating, though the specific measurement chosen varied.

Since the majority of measures in this dataset are ordinal, we would expect the polychoric method to be better suited at capturing the latent dependency structure of these symptoms. Therefore, interestingly, the strong preference of the polychoric method for UPDRS over BFM measures suggests that the diversity of XDP symptoms may be better captured by parkinsonism-related metrics rather than dystonia-related metrics, in this particular dataset at least. If forced to choose one of these two measurement scales over the other, this finding suggests that UPDRS might be of greater value for measuring XDP symptoms—an insight that is obscured by using Pearson correlations only.

5 Discussion

The standard approach to principal variables analysis depends on Pearson correlation, which is not well suited for non-Gaussian data. In diverse simulations, we found that PVA using Pearson correlation performs significantly worse than Spearman, Gaussian copula, and polychoric correlation for capturing the latent dependency structure of non-Gaussian data. Further, even though the Gaussian copula or polychoric correlation assume a latent multivariate Gaussian distribution, they still exhibited improved performance over Pearson when the latent distribution was multivariate t or multivariate Laplace, showing that their improved performance is robust to departures from the assumed model.

On the other hand, it is important to bear in mind that this study involved certain assumptions that will not always hold, and in practice, we recommend considering which type of correlation is most relevant to the application at hand. For instance, our simulation studies assumed that any ordinal variables were discretized approximations to continuous latent variables, and that our primary interest was in the dependence of those latent variables. However, if one is using PVA to select variables that will be included in a linear regression model—where the practical utility of the variables depends on their measurement type—it

Variable	Pearson	Spearman	Gaussian copula	Polychoric
BFM-D: Walking	1			
UPDRS 3.5: Left hand	2	5		
EAT-10, Q1: Swallowing issues led to weight loss	3		6	4
UPDRS 3.18: Constancy of rest tremor	4	3	3	
UPDRS 3.3: Neck rigidity	5	6		6
BFM-M: Left Leg	6			
BFM-D: Hygiene	7			
UPDRS 2.1: Speech	8		7	9
Reported age at onset	9	8		
UPDRS-3: Tremor (postural left)	10			
UPDRS 2.4: Eating		1	1	1
UPDRS 2.12: Walking and balance		2	2	2
EAT-10, Q2: Swallowing made going out difficult		4		
UPDRS 3.1: Speech		7		
UPDRS 3.13: Posture		9		
UPDRS 2.7: Handwriting		10	10	
UPDRS 3.4: Finger tapping (right)			4	5
UPDRS 3.17: Tremor (left upper extremity)			5	
BFM-M: Trunk			8	10
SVA Repeat Size			9	
UPDRS 3.17: Tremor (lip)				3
UPDRS 3.17: Tremor (right lower extremity)				7
UPDRS 1.2: Hallucinations				8

Table 1: First 10 measures chosen by each PVA method on the XDP data. The numbers 1-10 indicate the order in which the variables were chosen by each method. Blank entries are shown for variables not among the top 10 for each method.

may be better to choose variables that capture the structure of the observed data rather than the latent data. In a situation like this, Pearson correlation may be preferable to the presented alternatives, since the importance of variables in linear regression modeling depends on their observed linear correlations.

A direction for future work would be to consider other optimality criteria. Specifically, in this study we considered only the McCabe "variance-explained" criterion, in which the objective is to minimize the trace of the conditional covariance matrix of the omitted variables given the selected variables. Other optimization criteria, as well as PCA-based or clustering approaches to PVA, might yield different conclusions about the relative performance of Pearson, Spearman, Gaussian copula, and polychoric correlations for PVA. It would be interesting to study how the observed distribution of the data affects the performance of variable selection in a wider variety of PVA methods and algorithms.

References

P. Acuna, M. L. Supnet-Wells, N. A. Spencer, J. K. De Guzman, M. Russo, A. Hunt, C. Stephen, C. Go, S. Carr, N. G. Ganza, J. B. Lagarde, S. Begalan, T. Multhaupt-Buell, G. Aldykiewicz, L. Paul, L. Ozelius, D. C. Bragg, B. Perry, J. R. Green, J. W. Miller, and N. Sharma. Establishing a natural history of X-linked dystonia parkinsonism. *Brain Communications*, 5(3):fcad106, 4 2023. ISSN 2632-1297. doi: 10.1093/braincomms/fcad106.

T. Aneichyk, W. T. Hendriks, R. Yadav, D. Shin, D. Gao, C. A. Vaine, R. L. Collins, A. Domingo, B. Currall, A. Stortchevoi, T. Multhaupt-Buell, E. B. Penney, L. Cruz, J. Dhakal, H. Brand, C. Hanscom, C. Antolik, M. Dy, A. Ragavendran, J. Underwood, S. Cantsilieris, K. M. Munson, E. E. Eichler, P. Acuña, C. Go, R. D. G. Jamora, R. L. Rosales, D. M. Church, S. R. Williams, S. Garcia, C. Klein, U. Müller, K. C. Wilhelmsen, H. T. M. Timmers, Y. Sapir, B. J. Wainger, D. Henderson, N. Ito, N. Weisenfeld, D. Jaffe, N. Sharma, X. O. Breakefield, L. J. Ozelius, D. C. Bragg, and M. E. Talkowski. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell*, 172(5):897–909.e21, Feb 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.02.011.

E. M. L. Beale, M. G. Kendall, and D. W. Mann. The discarding of variables in multivariate analysis. *Biometrika*, 54(3-4):357–366, 1967.

D. C. Bragg, K. Mangkalaphiban, C. A. Vaine, N. J. Kulkarni, D. Shin, R. Yadav, J. Dhakal, M.-L. Ton, A. Cheng, C. T. Russo, M. Ang, P. Acuña, C. Go, T. N. Francoeur, T. Multhaupt-Buell, N. Ito, U. Müller, W. T. Hendriks, X. O. Breakefield, N. Sharma, and L. J. Ozelius. Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. *Proceedings of the National Academy of Sciences*, 114(51), Dec 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1712526114. URL `https://pnas.org/doi/full/10.1073/pnas.1712526114`.

M. J. Brusco. A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. *Computational Statistics & Data Analysis*, 77:38–53, 2014.

J. Cadima, J. Cerdeira, and M. Minhoto. Computational aspects of algorithms for variable selection in the context of principal components. *Computational Statistics & Data Analysis*, 47(2):225–236, 2004.

J. Choi, M. Peters, and R. O. Mueller. Correlational analysis of ordinal data: from Pearson's r to Bayesian polychoric correlation. *Asia Pacific Education Review*, 11(4):459–466, 12 2010. ISSN 1598-1037, 1876-407X.

J. Cumming and D. Wooff. Dimension reduction via principal variables. *Computational Statistics & Data Analysis*, 52(1):550–565, 2007.

Y. Dodge. *The Concise Encyclopedia of Statistics*. Springer reference. Springer New York : Imprint: Springer, New York, NY, 1st ed. 2008. edition, 2008. ISBN 0-387-32833-5.

I. T. Jolliffe. Discarding Variables in a Principal Component Analysis. I: Artificial Data. *Applied Statistics*, 21(2):160–173, 1972.

I. T. Jolliffe. Discarding Variables in a Principal Component Analysis. II: Real Data. *Applied Statistics*, 22(1):21–31, 1973.

S. Kotz, T. J. Kozubowski, and K. Podgórski. *The Laplace Distribution and Generalizations*. Springer Science+Business, Boston, MA, 2001. ISBN 978-1-4612-0173-1. URL `http://link.springer.com/10.1007/978-1-4612-0173-1`.

T. J. Kozubowski, K. Podgórski, and I. Rychlik. Multivariate generalized Laplace distribution and related random fields. *Journal of Multivariate Analysis*, 113:59–72, 1 2013. ISSN 0047259X.

L. V. Lee, F. M. Pascasio, F. D. Fuentes, and G. H. Viterbo. Torsion dystonia in Panay, Philippines. *Advances in Neurology*, 14:137–151, 1976. ISSN 0091-3952.

L. V. Lee, C. Rivera, R. A. Teleg, M. B. Dantes, P. M. D. Pasco, R. D. G. Jamora, J. Arancillo, R. F. Villareal-Jordan, R. L. Rosales, C. Demaisip, E. Maranon, O. Peralta, R. Borres, C. Tolentino, M. J. Monding, and S. Sarcia. The Unique Phenomenology of Sex-Linked Dystonia Parkinsonism (XDP, DYT3, "Lubag"). *International Journal of Neuroscience*, 121(sup1):3–11, 2 2011. ISSN 0020-7454, 1543-5245. doi: 10.3109/00207454.2010.526728.

S. Makino, R. Kaji, S. Ando, M. Tomizawa, K. Yasuno, S. Goto, S. Matsumoto, M. D. Tabuena, E. Maranon, M. Dantes, L. V. Lee, K. Ogasawara, I. Tooyama, H. Akatsu, M. Nishimura, and G. Tamiya. Reduced Neuron-Specific Expression of the TAF1 Gene Is Associated with X-Linked Dystonia-Parkinsonism. *The American Journal of Human Genetics*, 80(3):393–406, Mar 2007. ISSN 00029297. doi: 10.1086/512129.

G. P. McCabe. Principal Variables. *Technometrics*, 26(2):137–144, 1984.

U. Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460, Dec 1979. ISSN 0033-3123, 1860-0980. doi: 10.1007/BF02296207.

D. B. Rubin. Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics*, 4(1):87–94, 1986. ISSN 0735-0015. doi: 10.1080/07350015.1986.10509497.

P. K. Trivedi and D. M. Zimmer. Copula modeling: an introduction for practitioners. *Foundations and Trends in Econometrics*, 1(1):1–111, 2005. ISSN 1551-3076.

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

A Appendix

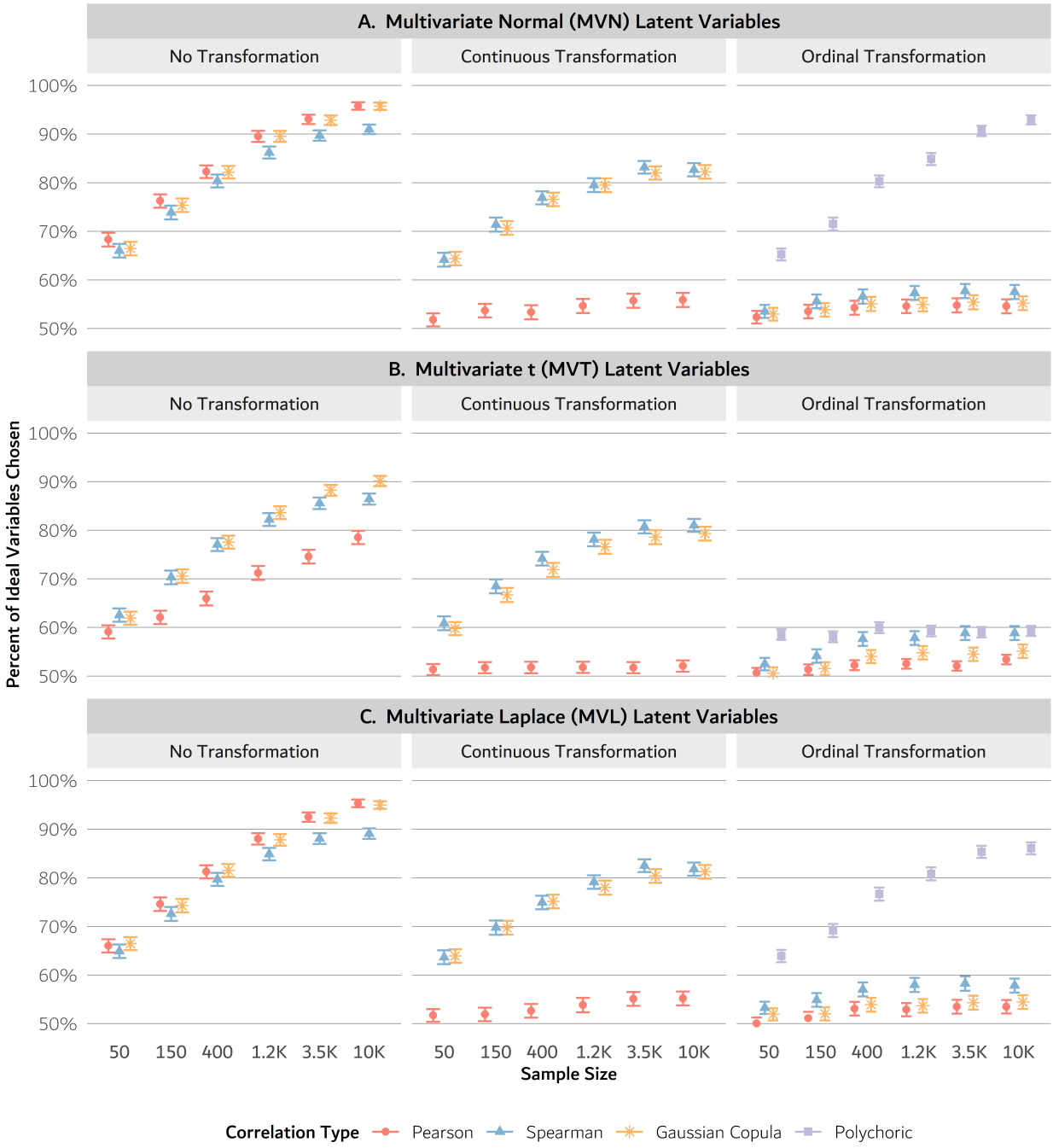

Figure A1: Proportion of latently ideal variables identified by PVA, varying the latent multivariate distribution. Value shown is the mean proportion over 1,000 replicate simulations, and error bars show ± two standard errors. The x-axis is on a logarithmic scale.

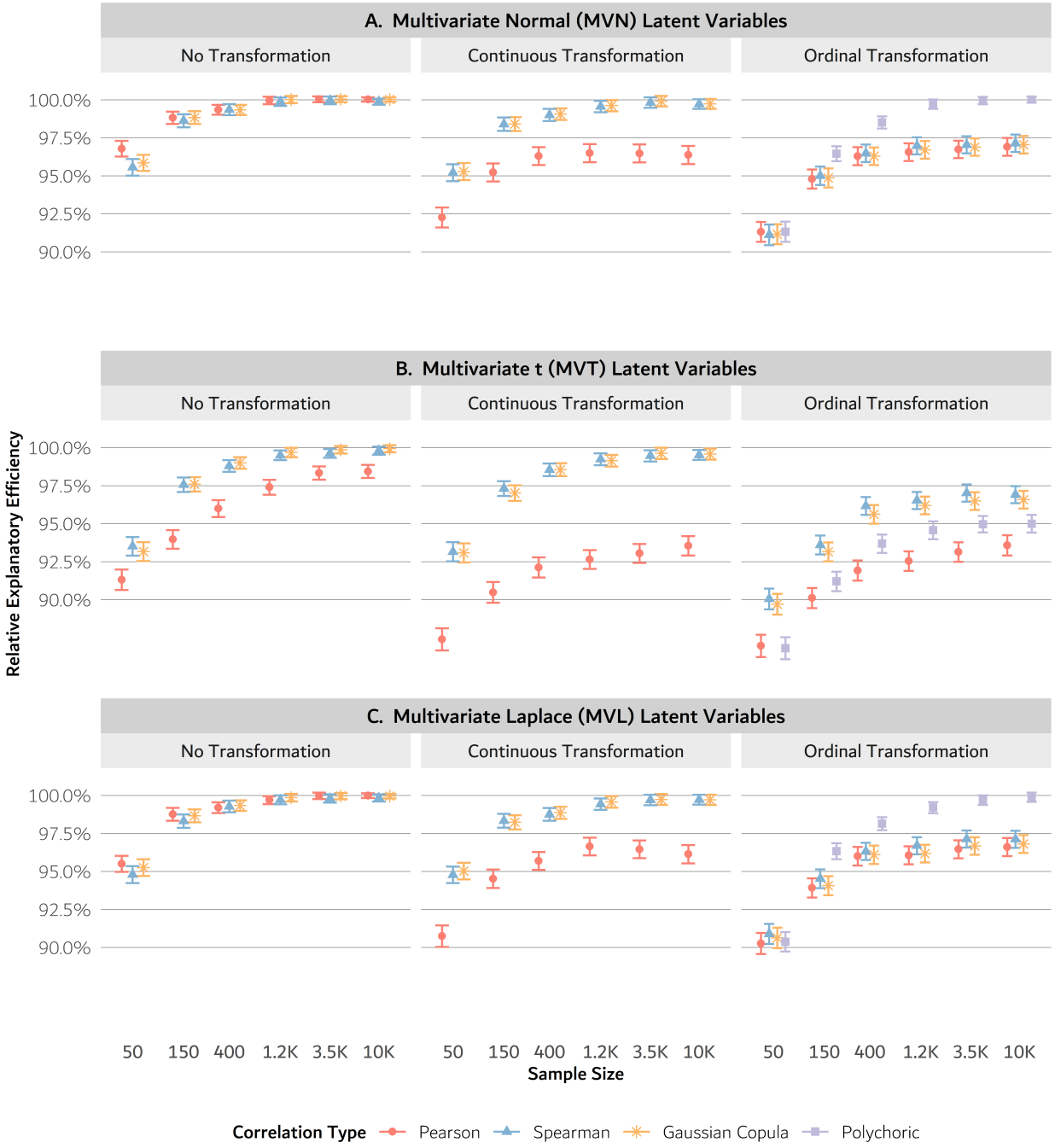

Figure A2: Relative explanatory efficiency of chosen variables to ideal variables, varying the latent multivariate distribution. The y-axis is the variance unexplained by the ideal variables divided by the variance unexplained by the selected variables, as a percentage. The value shown is the average over 1,000 simulations, and the error bars show $\pm$ two standard errors. Note that REE $> 100\%$ is mathematically possible.