Simulating High-Dimensional Multivariate Data

using the bigsimr R Package

Alfred G. Schissler Edward J. Bedrick Alexander D. Knudson
Tomasz J. Kozubowski Tin Nguyen Anna K. Panorska Juli Petereit
Walter W. Piegorsch Duc Tran

Abstract

It is critical to realistically simulate data when conducting Monte Carlo studies and methods. But measurements are often correlated and high dimensional in this era of big data, such as data obtained through high-throughput biomedical experiments. Due to computational complexity and a lack of user-friendly software available to simulate these massive multivariate constructions, researchers may resort to simulation designs that posit independence or perform arbitrary data transformations. This greatly diminishes insights into the empirical operating characteristics of any proposed methodology, such as false positive rates, statistical power, interval coverage, and robustness. This article introduces the `bigsimr` R package that provides a flexible, scaleable procedure to simulate high-dimensional random vectors with given marginal characteristics and dependency measures. We'll describe the functions included in the package, including multi-core and graphical-processing-unit accelerated algorithms to simulate random vectors, estimate correlation, and find close positive semi-definite matrices. Finally, we demonstrate the power of `bigsimr` by applying these functions to our motivating dataset — RNA-sequencing data obtained from breast cancer tumor samples with sample size $n = 1212$ patients and dimension $d = 1026$.

1 Introduction

Massive high-dimensional data sets are now commonplace in many areas of scientific inquiry. As new methods are developed for these data, a fundamental challenge lies in designing and conducting simulation studies to assess the operating characteristics of proposed methodology, such as false positive rates, statistical power, interval coverage, and robustness — often in comparison to existing methods. Further, efficient simulation empowers statistical computing strategies, such as the parametric bootstrap [Rizzo, 2007] to simulate from a hypothesized null model, providing inference in analytically challenging settings. Such Monte Carlo (MC)

techniques become difficult for high-dimensional data with the current existing algorithms and tools. This is particularly true when simulating massive *multivariate*, *non-normal* distributions, arising naturally in many fields of study.

As others have noted, it can be vexing to simulate dependent, non-normal/discrete data — even for low dimensional settings [Madsen and Birkes, 2013, Xiao and Zhou, 2019]. For continuous non–normal multivariate data, the well-known NORmal To Anything (NORTA) algorithm [Cario and Nelson, 1997] and other copula [Nelsen, 2007] approaches are well-studied with flexible, robust software available [Yan, 2007, Chen, 2001]. Yet these approaches do not scale in a timely fashion to high-dimensional problems [Li et al., 2019]. For discrete data, early simulation strategies had major flaws — such as failing to obtain the full range of possible dependencies (e.g., admitting only positive correlations Park et al. [1996]). While more recent approaches [Madsen and Birkes, 2013, Xiao, 2017, Barbiero and Ferrari, 2017] have largely remedied this issue for low-dimensional problems, the existing tools are not designed to scale to high dimensions.

Another central issue lies characterizing dependency between components in the high-dimensional random vector. The choice of correlation in practice usually relates to the eventual analytic goal and distributional assumptions of the data (e.g. non-normal, discrete, infinite support, etc). For normal data, the Pearson product-moment correlation describes the dependency perfectly. As we will see, however, simulating arbitrary random vectors that match a target Pearson correlation matrix exactly is computationally intense [Chen, 2001, Xiao, 2017]. On the other hand, an analyst may consider the use of nonparametric correlation measures to better characterize monotone, non-linear dependency, such as Spearman's ρ and Kendall's τ. Throughout, we'll emphasize matching these nonparametric dependency measures as our aim lies in modeling non-normal data and these rank-based measures possess invariance properties favorable in our proposed methodology.

With all this mind, we present a scaleable, flexible multivariate simulation algorithm. The crux of the method lies in the construction of a Gaussian copula, in the spirit of the NORTA procedure. Further, we introduce the `bigsimr` R package that provides parallelized, high-performance software implemented our NORTA-inspired algorithm. The algorithm design leverages useful properties of nonparametric correlation measures, namely invariance under monotone transformation and well-known closed form relationships between dependency measures for the multivariate normal (MVN) distribution.

The study proceeds by providing background information, including a description of our motivating example application — RNA-sequencing (RNA-seq) breast cancer data. Then we describe and justify our simulation methodology and related algorithms. Next, we detail an illustrative low-dimensional example of basic and advanced use of the `bigsimr` R package. Then we proceed with Monte Carlo studies under

various distributional assumptions to assess accuracy and scaleable. After the MC evaluations, we revisit our high-dimensional motivating example and employ our methods to commonplace statistical computing tasks. Finally, we'll discuss the method's utility, limitations, and future directions.

2 Background

The `bigsimr` `R` package provides multiple algorithms to work with high-dimensional multivariate data, but all these algorithms were originally designed to support a single task: to generate random vectors drawn from multivariate probability distribution with given marginal distributions and dependency metrics. Specifically, our goal is to efficiently simulate a large number, B, of random vectors $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top$ with **correlated** components and heterogeneous marginal distributions, described via cumulative distribution functions (CDFs) F_i, where d can be very large and still be computed in practically useful times.

When designing this methodology, we developed the following properties to guide our effort. We divide the properties into two categories: (1) basic properties (BP) and "scaleability" properties (SP). The BPs are adapted from an existing criteria due to Nikoloulopoulos [2013]. Our simulation strategy should allow:

- BP1: A wide range of dependences, allowing both positive and negative values, and, ideally, admitting the full range of possible values.
- BP2: Flexible dependence, meaning that the number of bivariate marginals can be equal to the number of dependence parameters.
- BP3: Flexible marginal modeling, generating heterogeneous data — possibly from differing probability families.

Moreover, the simulation method must **scale** to high dimensions:

- SP1: Procedure must scale to high dimensions, computable in a reasonable amount time.
- SP2: Procedure must scale to high dimensions while maintaining accuracy.

To fix ideas and provide examples applications enabled via `bigsimr`, the next section describes a motivating data set that originally inspired the authors' interest in developing this methodology.

2.1 Motivating example: RNA-seq data

Simulating high-dimensional, non-normal, correlated data motivates this work — in pursuit of modeling RNA-sequencing (RNA-seq) data [Wang et al., 2009, Conesa et al., 2016] derived from breast cancer patients. The RNA-seq data-generating process involves counting how often a particular messenger RNA (mRNA) is

Table 1: mRNA counts for three selected high-expressing genes from the first five observations of the BRCA data set.

RPL5	TXNIP	VIM
20283	13401	26883
18614	11365	28806
31378	5365	22221
37861	5873	26871
17902	12564	15985

expressed in a biological sample. RNA-seq platforms typically quantify the entire transcriptome in one experimental run, resulting in high-dimensional data. For human derived samples, this results in count data corresponding to over 20,000 genes (protein-coding genomic regions) or even over 77,000 isoforms when alternatively-spliced mRNA are counted. Importantly, due to inherent biological processes, gene expression data exhibits correlation — co-expression — across genes [Efron, 2007, Schissler et al., 2018].

We'll illustrate our methodology using the well-studied Breast Invasive Carcinoma (BRCA) data set housed in The Cancer Genome Atlas (TCGA; see Acknowledgments). For simplicity, we only consider high expressing genes. In turn, we begin by filtering to retain the top 5% highest- expressing genes (in terms of median expression) of the 20,501 gene measurements from $N = 1212$ patients' tumor samples, resulting in $d = 1026$ genes. This gives a massive number of pairwise dependencies among the marginals (specifically, 5.2582×10^5 correlation parameters). We further process the data to illustrate our methodology's flexible and robust simulation scheme, while better aligning with the actual data-generating process, by rounding Li and Dewey [2011] RNA-seq by Expectation Maximization (RSEM) values to counts. Table 1 displays counts for three selected high-expressing genes for the first five patients' breast tumor samples.

To help visualize the bivariate relationships for these three selected genes across all patients, Figure 1 plots the marginal distributions and estimated Spearman's correlations (see Equation (2) below).

2.2 Measures of dependency

In multivariate analysis, an analyst must select a metric to quantify dependency. The most widely-known is the Pearson (product-moment) correlation coefficient that describes the linear association between two random variables X and Y, and, it is given by

$$\rho_P(X, Y) = \frac{E(XY) - E(X)E(Y)}{[var(X)var(Y)]^{1/2}}. \tag{1}$$

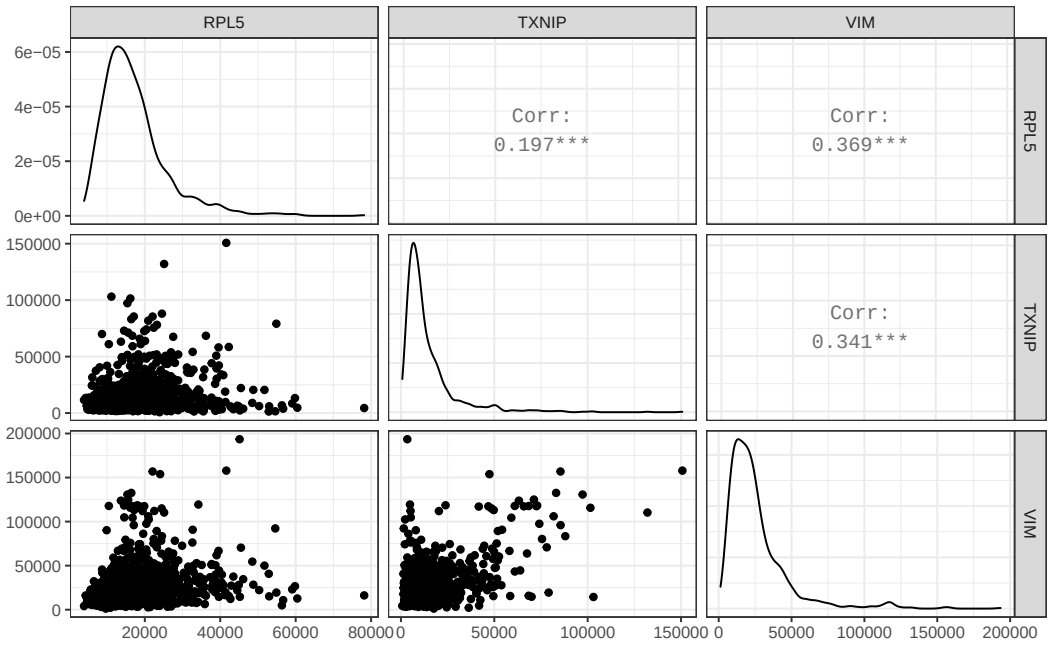

Figure 1: Marginal scatterplots, densities, and estimated pairwise Spearman's correlations for three example genes. The data possess heavy-right tails, are discrete, and have non-trivial intergene correlations. Modeling these data motivate our simulation methodology.

As Madsen and Birkes [2013] and Mari and Kotz [2001] discuss, for a bivariate normal (X, Y) random vector, the Pearson correlation completely describes the dependency between the components. For non-normal marginals with monotone correlation patterns, ρ_P suffers some drawbacks and may mislead or fail to capture important relationships (Mari and Kotz [2001]). Alternatively in these settings, analysts often prefer rank-based correlation measures to describe the degree of monotonic association.

Two nonparametric, rank-based measures common in practice are Spearman's correlation (denoted ρ_S) and Kendall's τ. Define

$$\rho_S(X, Y) = 3\left[P\left[(X_1 - X_2)(Y_1 - Y_3) > 0\right] - P\left[(X_1 - X_2)(Y_1 - Y_3) < 0\right]\right], \tag{2}$$

where $(X_1, Y_1) \overset{d}{=} (X, Y), X_2 \overset{d}{=} X, Y_3 \overset{d}{=} Y$ with X_2 and Y_3 are independent of one other and of (X_1, Y_1). Spearman's ρ_S has an appealing correspondence as the Pearson's correlation coefficient on *ranks* of the values, thereby captures nonlinear yet monotone relationships.

Kendall's τ, on the other hand, is the difference in probabilities of concordant and discordant pairs of observations (X_i, Y_i) and (X_j, Y_j). By concordance we mean that orderings have the same direction (e.g., if $X_i < X_j$, then $Y_i < Y_j$) and is determined by the ranks of the values, not the values themselves.

Both τ and ρ_S are **invariant to under monotone transformations** of the underlying random variates. As we will describe more fully in the Algorithms section, this property is essential to scaleable match rank-based correlations with speed (SP1) and accuracy (SP2).

Correspondence among Pearson, Spearman, τ correlations. This is no closed form, general correspondence among the rank-based measures and the Pearson correlation coefficient, as the marginal distributions F_i are intrinsic in their calculation. But for **bivariate normal vectors**, however, the correspondence is well-known:

$$\rho_P = sin\left(\tau \times \frac{\pi}{2}\right), \tag{3}$$

and similarly for Spearman's ρ [Kruskal, 1958],

$$\rho_P = 2 \times sin\left(\rho_S \times \frac{\pi}{6}\right). \tag{4}$$

These facts are also critical in our simulation algorithm to broaden the dependency measures supported by `bigsimr`, in a computationally effective manner.

Discrete marginal considerations. Spearman's correlation for discrete marginal suffers some issues due to the nonzero probability of ties (for example, the Spearman's correlation of a discrete-valued random variable X with itself could be less than 1; Madsen and Birkes [2013]). One remedy for this issue is to rescale Equation (2). For two random variables X, Y with probability mass functions (PMFs) or probability densities functions (PDFs) $p(x)$ and $q(y)$, respectively, define the rescaled Spearman's correlation as

$$\rho_{RS}(X,Y) = \frac{\rho_s(X,Y)}{\left[\left[1 - \sum_x p(x)^3\right]\left[1 - \sum_y q(y)^3\right]\right]^{1/2}}. \tag{5}$$

Note that with continuous marginals the rescaling returns ρ_S. For discrete marginals with large or infinite support, computing the adjustment factors $\sum_x p(x)^3$, $\sum_y p(y)^3$ over all large number of pairs becomes expensive (often violating desired property SP1). And further the infinite sums must be approximately for count-valued data and potentially violating desired property SP2.

Also, for a discrete random variable Y_i, some care must be taken to define the quantile function F_i^{-1}. Let

$$F_i^{-1} = inf\{y : F_i(y) \geq u\}. \tag{6}$$

Marginal-dependent bivariate correlation bounds. Given two marginal distributions, ρ_P is not free to vary the entire range of possible correlations $[-1, 1]$. The so-called *Frechet-Hoeffding bounds* are well-studied (for example, see Nelsen [2007] and Barbiero and Ferrari [2017]). This situation gives strict restraints on the possible correlations and cannot be overcome through algorithm design. In general, the bounds are given by

$$\rho_P^{max} = \rho_P\left(F_1^{-1}(U), F_2^{-1}(U)\right), \quad \rho_P^{min} = \rho_P\left(F_1^{-1}(U), F_2^{-1}(1-U)\right) \tag{7}$$

where U is a uniform random variable in $(0, 1)$, and F_1^{-1}, F_2^{-1} are the inverse CDF of random variables X_1 and X_2, respectively. For discrete random variables, define F^{-1} as in Equation (6).

2.3 Gaussian copulas

There is a strong connection of our simulation strategy to Gaussian **copulas** (see Nelsen [2007] for a technical introduction). A copula is a distribution function on $[0, 1]^d$ that describes a multivariate probability distribution with standard uniform marginals. This definition provides a powerful, natural way characterize joint probability structure. Consequently, the study of copulas is an important and active area of statistical theory and practice.

For any random vector $\mathbf{X} = (X_1, \ldots, X_d)$ with CDF F and marginal CDFs F_i there is a copula function $C(u_1, \ldots, u_d)$ so that

$$F(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d) = C(F_1(x_1), \ldots, F_d(x_d)), \;\; x_i \in \mathbb{R}, i = 1, \ldots, d.$$

A Gaussian copula is the case where all marginal CDFs F_i are the standard normal CDF, Φ. This representation corresponds to a multivariate normal distribution with standard normal marginal distributions and covariance matrix $\mathbf{R_P}$. But since the marginals are standardized to have unit variance, this $\mathbf{R_P}$ is a Pearson correlation matrix. If $F_{\mathbf{R}}$ is the CDF of such a multivariate normal distribution, then the corresponding Gaussian copula $C_{\mathbf{R}}$ is defined through

$$F_{\mathbf{R}}(x_1, \ldots, x_d) = C_{\mathbf{R}}(\Phi(x_1), \ldots, \Phi(x_d)), \tag{8}$$

where $\Phi(\cdot)$ is the standard normal CDF. Note that the copula $C_{\mathbf{R}}$ is the familar multivariate normal CDF of the random vector $(\Phi(X_1), \ldots, \Phi(X_d))$, where $(X_1, \ldots, X_d) \sim N_d(\mathbf{0}, \mathbf{R_P})$.

Sklar's Theorem [Sklar, 1959, Úbeda-Flores and Fernández-Sánchez, 2017] guarantees that given inverse CDFs F_i^{-1}s and a valid correlation matrix (within the Frechet bounds) a random vector can be obtained via transformations involving copula functions. For example, using Gaussian copulas, we can construct a random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ with $Y_i = F_i^{-1}(U_i)$, $i = 1, \ldots, d$, via $\mathbf{U} = (U_1, \ldots, U_d)$ viz $U_i = \Phi(X_i)$, $i = 1, \ldots, d$ povides $Y_i \sim F_i$, $\forall\, i$.

3 Algorithms

This section describes our methods involved in simulating a random vector $\mathbf{Y}$ with Y_i components for $i = 1, 2, \ldots, d$. Each Y_i has a specified marginal CDF F_i and its inverse F_i^{-1}. To characterize dependency, every pair (Y_i, Y_j) has a given Pearson correlation ρ_P, Spearman correlation ρ_S, and/or Kendall's τ. The method is best understand as a **high-performance Gaussian copula** (Equation (8)) providing a high-dimensional NORTA-inspired algorithm.

3.1 NORmal To Anything (NORTA)

The well-known NORTA algorithm [Cario and Nelson, 1997] can be used simulate a random vector $\mathbf{Y}$ with variance-covariance matrix $\Sigma_{\mathbf{Y}}$. Specifically, the NORTA algorithm follows like this:

1. Simulate a random vector $\mathbf{Z}$ with d **independent** and **identical** standard normal components.
2. Determine the input matrix $\Sigma_{\mathbf{Z}}$ that corresponds with the specified output $\Sigma_{\mathbf{Y}}$.
3. Produce a Cholesky factor M of $\Sigma_{\mathbf{Z}}$ so that $MM' = \Sigma_{\mathbf{Z}}$.
4. Set X by $X \leftarrow MZ$.
5. Return Y where $Y_i \leftarrow F_{Y_i}^{-1}[\Phi(X_i)],\ i = 1, 2, ..., d$.

With modern parallelized computing, steps 1, 3, 4, 5 are readily implemented as high-performance, multi-core and/or graphical-processing-unit (GPU) accelerated, algorithms — providing the fast scaleability using readily-available hardware.

Matching specified Pearson correlation coefficients exactly (step 2 above), however, is problematic. In general, there is no closed form correspondence between the components of the input $\Sigma_{\mathbf{Z}}$ and target $\Sigma_{\mathbf{Y}}$. Matching the correlations involves evaluating or approximating $\binom{d}{2}$ integrals of the form $EY_iY_j = \int\int y_i y_j f_X(F_i^{-1}(\Phi(z_i)), F_j^{-1}(\Phi(z_j))dy_i dy_j$, for $i, j = 1, 2, \ldots, d,\ i \neq j$. For high-dimensional data, these evaluations are often too costly to enable feasible simulation studies. For low-dimensional problems, methods and tools exist to match Pearson correlations precisely, see Chen [2001]; Xiao [2017]; Madsen and Birkes [2013] and the publicly available `nortaRA` R package.

To maintain scaleability (SP1), our solution is to essentially avoid this complication in Pearson matching. Since our goal is to simulate non-normal marginals, we greatly prefer the use of rank-based measures ρ_S and τ from a modeling standpoint. Further, ρ_S and τ's invariance under monotone transformation (see Background), preserves the correlation coefficients through steps 3, 4, and 5 in the NORTA algorithm above. This eliminates the need for computing the $\binom{d}{2}$ integrals to match exactly (nothing is for free, however, as discussed in below in Section 3.2).

Despite all this, if one does desire to characterize dependency using Pearson correlations, simply using the target Pearson correlation matrix as the initial conditions to our proposed algorithm will lead to approximate matching in the resultant distribution [Song, 2000] in many practical applications. The quality of this approximation depends on the setting, but in practice, for high-dimensional count data we find the accuracy to be adequate. Later, we'll study the robustness of our method to this limitation in selected Monte Carlo evaluations.

3.2 Random vector generation via `bigsimr::rvec`

Now we describe `bigsimr::rvec`, our algorithm to generate random vectors. It mirrors the classical NORTA algorithm above with some modifications for rank-based dependency matching:

1. Pre-processing for nonparametric dependency matching.
   (i) Convert from either $\mathbf{R}_{\mathbf{Spearman}}$ or $\mathbf{R}_{\mathbf{Kendall}}$ into the corresponding MVN input correlation $\mathbf{R}_{\mathbf{Pearson}}$.
   (ii) Check that $\mathbf{R}_{\mathbf{Pearson}}$ is semi-positive definite.
   (iii) If not compute a close semi-positive definite correlation matrix $\tilde{\mathbf{R}}_{\mathbf{Pearson}}$.
2. Gaussian copula construction.
   (i) Generate $\mathbf{X} = (X_1, \dots, X_d) \sim N_d(\mathbf{0}, \mathbf{R}_{\mathbf{Pearson}})$.
   (ii) Transform $\mathbf{X}$ to $\mathbf{U} = (U_1, \dots, U_d)$ viz $U_i = \Phi(X_i)$, $i = 1, \dots, d$.
3. Quantile evaluations.
   (i) Return $\mathbf{Y} = (Y_1, \dots, Y_d)$, where $Y_i = F_i^{-1}(U_i)$, $i = 1, \dots, d$;

The pre-processing (Step 1) takes advantage of the closed-form relationships between ρ_S and τ with

ρ_P for bivariate normal random variables via Equations (4) or (3), respectively (implemented as `bigsimr::cor_covert`).

A complication often arises at this stage: the resultant matrix may become indefinite, either through numerical error or naturally occurring in the correlation conversion. Researchers working in multivariate computation frequently encounter such difficulties and need to find a close positive (semi-)definite matrix. A widely-available routine for this task in `R` is called `matrix::nearPD`, though it is not suitable for high dimensions. To overcome this issue, we've developed `bigsimr::nearPSD`, a quadractically-convergent Newton method for finding the nearest correlation matrix, developed by Qi and Sun [2006]. We hope that this routine could be useful in many applications aside from our primary goal of random vector generation.

Once the target margins and algorithm inputs are determined, steps 2 and 3 are essentially a NORTA algorithm with modern high-performance computing implementations. Specifically, step 2i uses either an efficient multi-core multivariate normal simulator (the `R` package `mvnfast`[Fasiolo, 2016]) or a using Google's JAX python library NumPy for graphical-processing-unit (GPU) acceleration of the Cholesky factorization and matrix multiplication (steps 3, 4 in the NORTA algorithm in the preceding section).

3.3 A note on NORTA in higher dimensions

Sklar's theorem provides a useful characterization of multivariate distributions through copulas. Yet the choice of copula-based simulation algorithm affects which joint distributions may be simulated. Even in low-dimensional spaces (e.g., $d = 3$), there exists valid multivariate distributions with *feasible* Pearson correlation matrices that NORTA cannot match exactly [Ted Li and Hammond, 1975]. This occurs when the bivariate transformations are applied to find the input correlation matrix yet when combined the resultant matrix is indefinite. These situations do occur, even using exact analytic calculations. Such problematic target correlation matrices are termed *NORTA defective*.

Ghosh and Henderson [2002] conducted a Monte Carlo study to estimate the probability of NORTA defective matrices while increasing the dimension d. They found at what is now considered low-to-moderate dimensions ($d \approx 20$) that almost *all* feasible matrices are NORTA defective. This stems from the concentration of measure near the boundary of the space of all possible correlation matrices as dimension increases. Unfortunately, it is precisely near this boundary that NORTA defective matrices reside.

There is hope, however, as Ghosh and Henderson [2002] also show that by augmenting the NORTA procedure by replacing the indefinite input correlation matrix with a close proxy will give approximate matching to the target — with adequate performance for moderate d. This provides evidence that our nearest positive

semi-definite (PSD) augmented approach will maintain reasonable accuracy if our input matching scheme returns an indefinite matrix, at least for the rank-based matching scheme described above.

4 The `bigsimr` R package

The `bigsimr` package is a high-performance implementation of the proposed random vector generation algorithm and associated functions (see Section 3 for details). When designing `bigsimr`, we aimed to conveniently provide parallelized computation, through multi-core and GPU acceleration, while allowing advanced users to customize and automate workflows.

This subsections below describe the basic use of `bigsimr`, by stepping through a low-dimensional (2D) simulation workflow for the often-used example data set `airquality`. This workflow proceeds from data, to estimation, simulation configuration, random vector generation, and result visualization. For this low-dimensional setting, we compute using a single central processing unit (CPU) as the overhead in forking the tasks to multiple cores outweighs the computational gains. Then we transition to advanced use where we briefly describe some of the high-performance features and syntax. The section concludes with a short description of how to use `bigsimr` on a computer clusters through slurm scheduling via the `rslurm` package.

4.1 Basic use illustrated through a minimal example

We'll demonstrate the basic use and syntax of `bigsimr` through an example workflow applied to the New York air quality data set (`airquality`) included in the R `datasets` package. First, we load the `bigsimr` library and a few other convenient data science packages, including the syntactically-elegant `tidyverse` suite of `R` packages.

```
library(bigsimr)
library(tidyverse)
library(patchwork)
```

For simplicity and to provide a minimal working example, we'll consider bivariate simulation of temperature, in degrees Fahrenheit, and ozone level, in parts per billion.

```
df <- airquality %>%
  select(Temp, Ozone) %>%
  drop_na()
```

```
Rows: 116
```

```
Columns: 2
$ Temp  <int> 67, 72, 74, 62, 66, 65, 59, 61, 74, 69, 66, 68, 58, 64, 66, 5...
$ Ozone <int> 41, 36, 12, 18, 28, 23, 19, 8, 7, 16, 11, 14, 18, 14, 34, 6, ...
```

Figure 2 visualizes the bivariate relationship between Ozone and Temperature. We aim to simulate random two-component vectors mimicking this structure. The margins are not normally distributed, particularly the ozone level exhibits a strong positive skew.

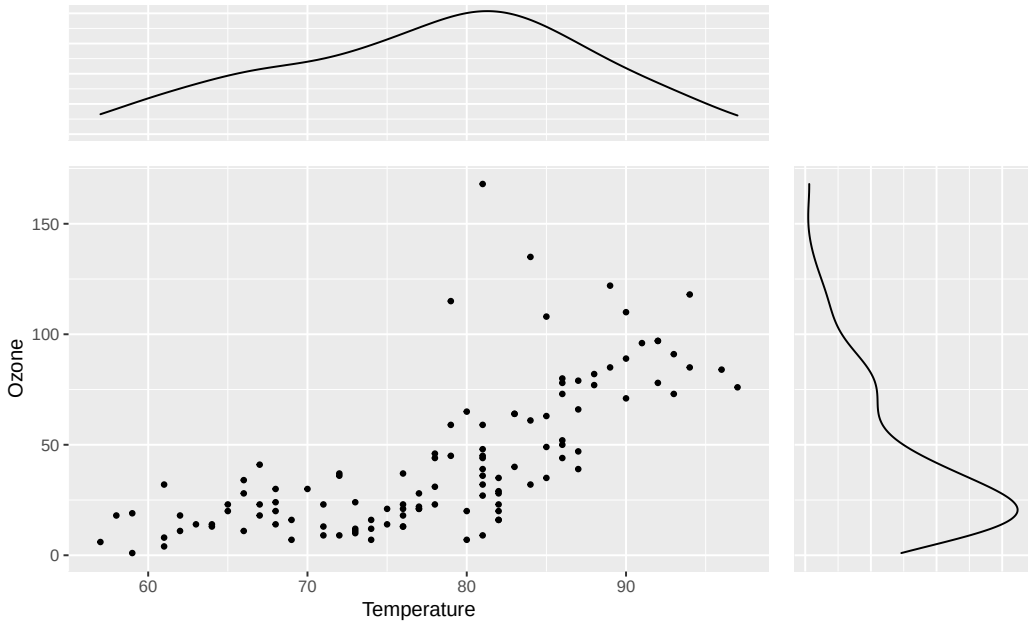

Figure 2: Bivariate scatterplot of Ozone vs. Temperature with estimated marginal densities. The data are left skewed tails and appear to be correlation.

Next, we specify the marginal distributions and correlation coefficient (both its type and magnitude). Here the analyst is free to be creative. For this example, we will not take up goodness-of-fit considerations to determine the marginal distributions. But it seems sensible without domain knowledge to estimate these quantities from the data and `bigsimr` contains fast functions designed for this task.

Specifying marginal distributions. Based on the estimated densities in Figure 2, we'll assume `Temp` is normally distributed and `Ozone` is log-normally distributed since its values are positive and skewed. We'll use the well-known, unbiased estimators for the normal distribution's parameters and maximum likelihood estimators for the log-normal parameters:

```
df %>%
  select(Temp) %>%
```

```
  summarise_all(.funs = c(mean = mean, sd = sd))
   mean    sd
1 77.87 9.485
```

```
mle_mean <- function(x) mean(log(x))
mle_sd <- function(x) mean( sqrt( (log(x) - mean(log(x)))^2 ) )


df %>%
  select(Ozone) %>%
  summarise_all(.funs = c(meanlog = mle_mean, sdlog = mle_sd))
  meanlog  sdlog
1   3.419 0.6967
```

Next, we'll configure the input marginals for later input into `bigsimr::rvec`. The marginal distributions are specifying using `R`'s special `alist` function. This allows one to enter the distributions without evaluating anything (yet).

```
margins <- alist(
    qnorm(mean = 77.871, sd = 9.4855),
    qlnorm(meanlog = 3.419, sdlog = 0.6967),
)
```

Notice that we use the *quantile* function for the marginals, as that is how the marginal distributions F_i enter into the `bigsimr::rvec` algorithm. This implementation strategy supports all `R base` probability distributions. And allows flexible extensions using other `R` packages that adhere to conventions, such as `extraDistr`. Further, by using `alist`, users can specify their own custom distributions (see below in creating custom margins).

It is a bit inconvenient to have to fill in the parameter values manually each time, so we provide a convenience function `mlist` which behaves similarly to `alist`, except that it will evaluate the right hand side of argument values within the list. This is intended to help when scaling up your code to high dimensions when many marginals to specify.

```
margins <- mlist(
  qnorm(mean = mean(df$Temp), sd = sd(df$Temp)),
  qlnorm(meanlog = mle_mean(df$Ozone), sdlog = mle_sd(df$Ozone))
```

```
)
margins
[[1]]
qnorm(mean = 77.8706896551724, sd = 9.48548563759966)


[[2]]
qlnorm(meanlog = 3.41851510081201, sdlog = 0.696668863646896)
```

Specifying correlation. As mentioned, the user must decide how to describe correlation, based on the particulars of the problem. For non-normal data and for improved simulation accuracy in our scheme, we advocate the use of rank-based correlations Spearman's ρ_S and Kendall's τ. But we also support approximate Pearson correlation coefficient matching, while cautioning the user to check the performance for their parametric multivariate model (see Monte Carlo evaluations for evaluation strategies and guidance). To aid in correlation specification, and estimation in general, we provide a high-performance function `bigsimr::cor_fast` which estimates Pearson, Spearman, or Kendall correlation using the fastest methods available. (Anyone who has tried estimating Kendall's τ using `stats::cor` can attest that the routine does not scale to even moderate dimensions). Notably, these estimation methods are the standard approaches, not designed specifically designed for high-dimensional correlation estimation (see Conclusion and Discussion for more on this).

```
type <- 'spearman'
(rho <- cor_fast(df, method = "spearman"))
       Temp Ozone
Temp  1.000 0.774
Ozone 0.774 1.000
```

Checking the theoretical correlation bounds As discussed in Section 2, given a pair of marginal distributions the possible correlations are not free to vary between $[-1, 1]$. To ensure that the simulation is not configured to impossible settings, we provide the `bigsimr::cor_bounds` function provides MC estimated theoretical lower and upper bounds (using the Generate, Sort, and Correlate algorithm of Demirtas and Hedeker [2011]).

```
cor_bounds(margins = margins, type = type)
$lower
     [,1] [,2]
[1,]    1   -1
[2,]   -1    1
```

```
$upper
     [,1] [,2]
[1,]    1    1
[2,]    1    1
```

Since our estimated Spearman correlation $\widehat{\rho}_S$ is within the theoretical bounds, the correlation is valid as input to `bigsimr:rvec`. On the other hand, the bounds on the Pearson correlation coefficient ρ_P between these margins is

```
cor_bounds(margins = margins, type = 'pearson', reps=1e6)
$lower
       [,1]   [,2]
[1,]  1.000 -0.881
[2,] -0.881  1.000

$upper
       [,1]   [,2]
[1,] 1.0000 0.8806
[2,] 0.8806 1.0000
```

Simulating random vectors. Finally, we arrive at the main function of `bigsimr`, `rvec`. Let's now simulate $B = 10,000$ random vectors from the assumed joint distribution of Ozone levels and Temp.

```
x <- rvec(10000, rho, margins, type)
df_sim <- as.data.frame(x)
colnames(df_sim) <- colnames(df)
```

Figure 3 plots the 10,000 simulated points.

4.2 Advanced use

Creating a custom marginal distribution. Because `bigsimr` uses an `alist` to store the margins, any probability distribution with a well-defined inverse CDF can be used including custom marginal distributions not provided in base `R`. Therefore, to specify a marginal distribution absent from an existing package, the user needs to provide a closed-form expression to form the corresponding quantile function.

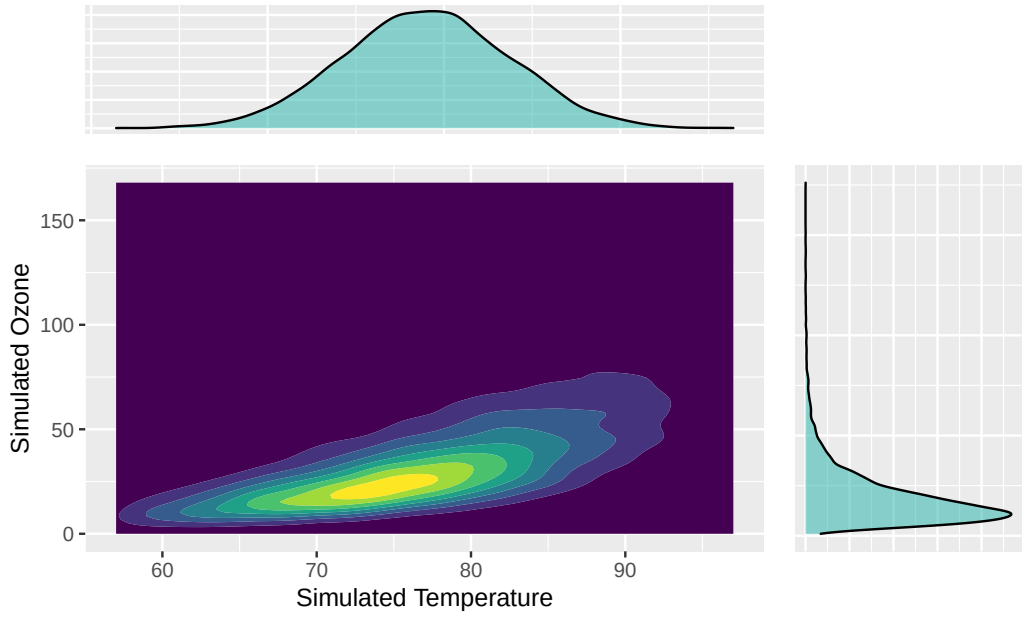

Figure 3: Contour plot and marginal densities for the simulated bivariate distribution of Air Quality Temperatures and Ozone levels. The simulated points mimic the observed data with respect to both the marginal characteristics and bivariate association.

It is important to follow `R`'s naming convention for probability distributions. Users should prefix distributions with `r` and `q` for *random* and *quantile* respectively. Only the quantile function is necessary to simulate random vectors, but to compute the theoretical correlation bounds, it is required to supply univariate random number generator as well.

For an example, let's provide a custom *Pareto* distribution for use with `bigsimr`. The Pareto CDF is

$$F(x) = 1 - \left(\frac{x_m}{x}\right)^{\alpha}$$

for scale $x_m > 0$ and shape $\alpha > 0$, with support $x \in [x_m, \infty)$. From the CDF, we compute the inverse CDF

$$F^{-1}(p) = \frac{x_m}{(1-p)^{1/\alpha}}$$

Next we define the Pareto quantile function in R:

```
qpareto <- function(p, scale, shape) {
  scale / (1 - p)^(1/shape)
}
```

Writing random number generating function for our target marginal can be accomplished by calling the quantile function on a uniformly distributed random variable (the inverse transform method Rizzo [2007]).

```
rpareto <- function(n, scale, shape) {
  qpareto(runif(n), scale, shape)
}
```

Now with the `qpareto` and `rpareto` functions, we can use the distribution in `bigsimr` just like the other built-in distributions.

```
margins <- alist(
  qnorm(mean = 3.14, sd = 0.1),
  qbeta(shape1 = 1, shape2 = 4),
  qnbinom(size = 10, prob = 0.75),
  qpareto(scale = 1.11, shape = 5.55)
)
cor_bounds(margins, "pearson")
rho <- cor_randPD(4)
x <- rvec(10, rho, margins)
```

Using `bigsimr` on a computing cluster via `rslurm`. Though `bigsimr` runs quickly, at large d users may want to run jobs on a shared computing server. The R package `rslurm` makes it easy to run embarrassingly large parallel `rvec` calls. This example assumes that `bigsimr` is installed on a system with a slurm scheduler installed. A single run of `bigsimr:rvec` using `rslurm` can be code as:

```
library(rslurm)
sjob <- slurm_call(rvec, jobname = 'rvec',
                   alist(n=1e6,
                         rho = rho,
                        margins = margins,
                        type = "spearman"),
                   submit = TRUE)
```

Now, let's show off the real power of combining `bigsimr` and `rslurm` by simulating many correlation structures. The `rslurm::slumr_map` syntax mirrors the `base::lapply` and `purrr::map` functions.

```
library(rslurm)
simReps <- 100
rhoList <- replicate( n = simReps, bigsimr::cor_randPSD(d = length(margins)),
                      simplify=FALSE )
sjob <- slurm_map(x = rhoList,
                  f = rvec,
                  jobname = 'rvecMap',
                  n=1e6,
                  margins = margins,
                  type = "spearman",
                  nodes = 4,
                  cpus_per_node = 4,
                  cores = 4,
                  submit = TRUE)
```

On a cluster carrying 24 nodes with 48 threads, these 100 jobs completed in about a minute.

5 Monte Carlo evaluations

Before applying our methodology to real data simulation, we conduct several Monte Carlo studies to investigate method performance. Since marginal parameter matching in our scheme is essentially a sequence of univariate inverse probability transforms, the challenging aspects are the accuracy of dependency matching and computational efficiency at high dimensions. To evaluate our methods in those respects, we design the following numerical experiments to first assess accuracy in match dependency parameters in bivariate simulations and then time the procedure in increasingly large dimension d.

5.1 Bivariate experiments

We select bivariate simulation configurations to ultimately simulate discrete-valued RNA-seq data and, so, proceed in increasing complexity, leading to the model in our motivating application in Section 6. We begin with empirically evaluating the dependency matching across all three supported correlations — Pearson's, Spearman's, and Kendall's —- in identical, bivariate marginal configurations. For each pair of identical margins, we vary the correlations across the entire possible range of values to evaluate the simulation's ability to obtain the theoretic bounds. The simulations progress from bivariate normal, to bivariate gamma

(non-normal yet continuous), and bivariate negative binomial (mimicking RNA-seq counts).

Table 2 lists our identical-marginal, bivariate simulation configurations. We increase the simulate replicates B to check that our results converge to the target correlations and gauge statistical efficiency. We select distributions beginning with a standard multivariate normal (MVN) as we expect the performance to be exact (up to MC error) for all correlation types. Then, we select a non-symmetric continuous distribution: a standard (rate =1) two-component multivariate gamma (MVG). Finally, we select distributions and marginal parameter values that are motivated by our RNA-seq data, namely values proximal to probabilities and sizes estimated from the data (see Example applications for our motivating data for estimation details). Thus we arrive at a multivariate negative binomial (MVNB) $p_1 = p_2 = 3 \times 10^{-4}, r_1 = r_2 = 4, \rho$).

Table 2: Identical margin, bivariate simulation configurations to evaluate correlation matching accuracy and efficiency.

Simulation Reps (B)	Correlation Types	Identical-margin 2D distribution
1000	Pearson (ρ_P)	$\mathbf{Y} \sim MVN(\mu = 0, \sigma = 1, \rho)$
$10,000$	Spearman (ρ_S)	$\mathbf{Y} \sim MVG(shape = 10, rate = 1, \rho)$
$100,000$	Kendall (τ)	$\mathbf{Y} \sim MVNB(p = 3 \times 10^{-4}, r = 4, \rho)$

For each of the unique 9 simulation configurations above, we estimate the correlation bounds and vary ρ along a sequence of 100 points evenly placed within the bounds (minus an adjustment factor of $\epsilon = 0.01$ to handle numeric issues when the bound is specified exactly).

Figure 4 displays the aggregated bivariate simulation results. Table 3 contains the mean absolute error (MAE) in reproducing the desired dependency measures for the three bivariate scenarios. Taken together, the studies show our methodology is generally accuracy across the entire range of possible correlation values for the rank-based dependency measures, at least in these limited simulation settings for the rank-based correlations. For the two non-normal bivariate marginals, the Pearson correlation matching is approximate. For discrete margins, matching the dependency measures was somewhat less accurate, even for the rank-based metrics, and particularly inaccurate near the lower bound of Pearson correlations.

The accuracy appears adequate for many applications. In practice, we recommend users to always evaluate the accuracy for their application, using methods similar to those presented above. See Discussion for future directions for fast Pearson matching and discrete-specific modifications.

Table 3: Average abolute error in matching the target dependency across the entire range of possible correlations for each bivariate marginal.

No. of random vectors	Correlation type	Distribution	Mean abs. error
1000	Pearson	MVN	0.0151
1000	Pearson	MVG	0.0217
1000	Pearson	MVNB	0.0326
1000	Spearman	MVN	0.0182
1000	Spearman	MVG	0.0169
1000	Spearman	MVNB	0.0159
1000	Kendall	MVN	0.0111
1000	Kendall	MVG	0.0122
1000	Kendall	MVNB	0.0114
10000	Pearson	MVN	0.0055
10000	Pearson	MVG	0.0120
10000	Pearson	MVNB	0.0246
10000	Spearman	MVN	0.0056
10000	Spearman	MVG	0.0056
10000	Spearman	MVNB	0.0050
10000	Kendall	MVN	0.0040
10000	Kendall	MVG	0.0033
10000	Kendall	MVNB	0.0031
1e+05	Pearson	MVN	0.0017
1e+05	Pearson	MVG	0.0102
1e+05	Pearson	MVNB	0.0227
1e+05	Spearman	MVN	0.0017
1e+05	Spearman	MVG	0.0018
1e+05	Spearman	MVNB	0.0018
1e+05	Kendall	MVN	0.0012
1e+05	Kendall	MVG	0.0011
1e+05	Kendall	MVNB	0.0010

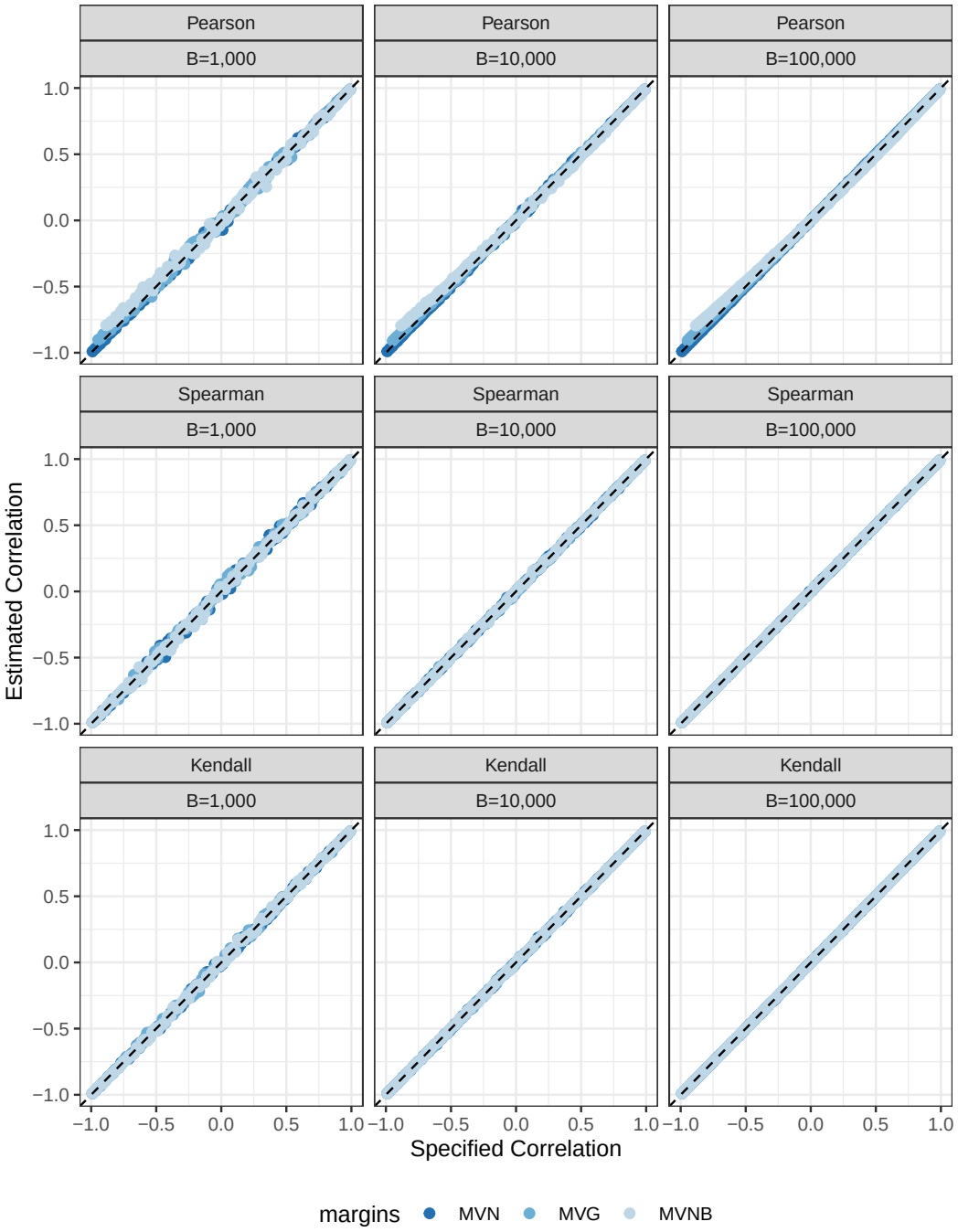

Figure 4: Bivariate simulations match specified correlations. The horizontal axis plots the specified target correlations across the entire range of possible correlations for each bivariate margin. Normal margins are plotted in dark blue, gamma in medium blue, and negative binomial in light blue. As the number of simulated vectors B increases from left to right, the variation in estimated correlations (vertical axis) decreases. The dashed line indicates equality between the specified and estimated correlations. Only the normal margins match the Pearson correlations (top row) exactly across the entire range of correlations. The two non-normal bivariate random vectors experience attenuation (downward bias) for extreme Pearson correlations (due to limitations in our algorithm) and more restriction (due to the Frechet bounds). The rank-based correlations (bottom two rows) are matched exactly for all margins and obtain the full range of possible dependencies.

5.2 Scale up to High Dimensions

With information of our method's accuracy from a low-dimensional perspective, we now turn to assessing whether the `bigsimr` can scale to larger dimensional problems. Specifically, we seek evidence to determine whether the method can scale to higher dimensions in a practical time. Using our motivating RNA-seq data, described in Background, we filtered the original 20,501 genes to the high-expressing genes at increasing percentiles, $1, 5, 10, 15, 20, 25\%$, to obtain $d = \{206, 1026, 2051, 3076, 4101, 5127\}$ marginals and $\binom{d}{2}$ pairwise correlations at each setting. For example, for $d = 5127$ there are $13,140,501$ correlation coefficients. We estimated the marginal negative binomial parameters and the correlation coefficients from the RNA-seq data to seed our simulations. (See Example applications for our motivating data for a detailed description of estimation).

Figure 5 displays computation times using various high-performance settings (1 central processing unit, CPU-1, versus twenty CPUs, CPU-20; with and without GPU acceleration, GPU-1; GPU-20) to produce $B = 10,000$ random vectors. The Pearson simulations are much faster since the correlation conversion steps are avoided (pre-processing step; see Algorithms), but as we know from above the accuracy will suffer slightly, especially near the negative boundary of the possible correlations. Matching Spearman's correlation at larger d gets costly if one wants to produce $B = 10,000$ random vectors at many different simulation settings, but the conversion steps need only be computed once (see MC evaluation of correlation estimation efficiency for an example of this strategy).

6 RNA-seq data example applications

This section demonstrates how to simulate multivariate data using `bigsimr`, aiming to replicate the structure of high-dimensional dependent count data. In an illustration of our proposed methodology applied to real data, we seek to simulate RNA-sequencing data by producing simulated random vectors with assumed marginal distributions with estimated parameters that mimic the observed data and its generating process. We seek scaleable realistic multivariate simulation to enable Monte Carlo (MC) methods for these data. Modeling RNA-seq using multivariate probabilities distributions is natural as inter-gene correlation is an inherent part of biological processes. Yet many models do not account for this, leading to major disruptions to the operating characteristics of statistical estimation, testing, and prediction. See Efron [2012] for a detailed discussion with related methods and see Wu and Smyth [2012], Schissler et al. [2018], Schissler et al. [2019] for applied examples. The following subsections apply `bigsimr`'s methods to real RNA-seq data, including replicated an estimated parametric structure, MC probability estimation, and MC evaluation of

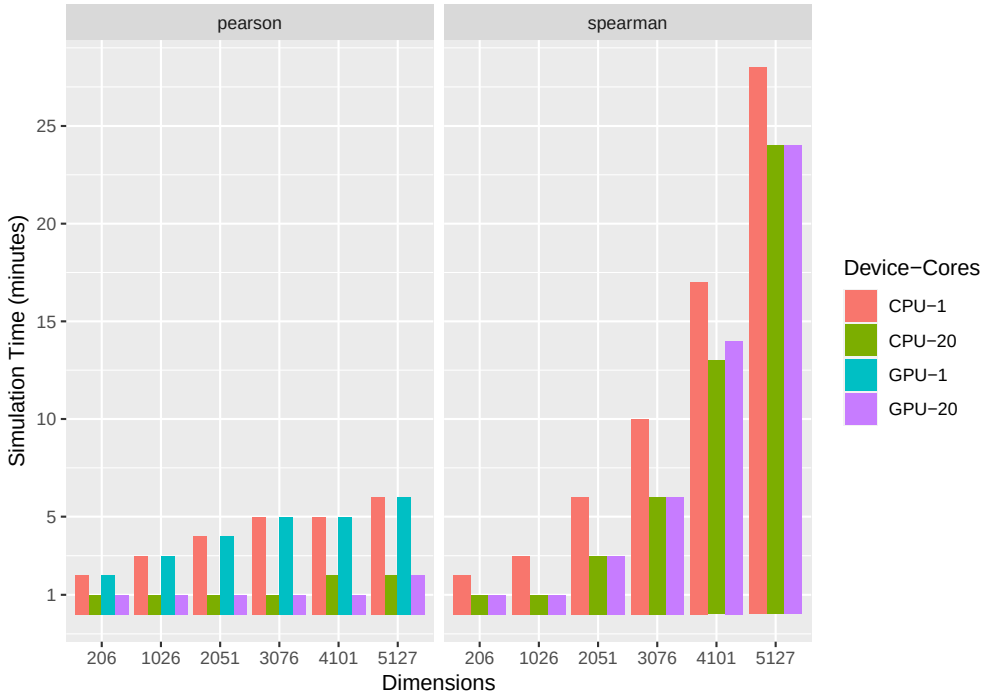

Figure 5: Computation times as d increases. We filter to the top 1, 5, 10, 15, 20, 25% expressing genes (in terms of median expression.)

correlation estimation efficiency.

6.1 Simulating High-Dimensional RNA-seq data

Our first goal is to replicate the structure of the TCGA BRCA RNA-seq data set (see Background). Ultimately, we will simulate $B = 10,000$ random vectors $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top$ with $d = 1026$. Often researchers posit a negative binomial (NB) model as RNA-seq counts are often over-dispersed that a Poisson model would suggest. All d selected genes exhibit over-dispersion (data not shown) and, so, we proceed to estimate the NB parameters $(r_i, p_i), i = 1, \ldots, d$ to determine the target marginal PMFs f_i (via method of moments). To complete specification of the simulation algorithm inputs, we'll estimate the Spearman correlation matrix $\mathbf{R}_{Spearman}$ to characterize dependency.

With our goal in mind, we first estimate the desired correlation matrix using the fast implementation provided by `bigsimr`:

```
## Estimate Spearman's correlation on the count data
corType <- 'spearman'
system.time( nb_Rho <- bigsimr::cor_fast( brca, method = corType ) )
   user  system elapsed
```

```
  0.621   0.000   0.622
```

Next, we estimate the marginal parameters. We use method of moments (MoM) to estimate the marginal parameters for the multivariate negative binomial model. The marginal distributions are from the same probability family (NB) yet are heterogeneous in terms of the parameters probability and size (p_i, n_i) for $i, \ldots, d$. The functions below help achieve this estimation and specifying the inputs for use in `bigsimr::rvec`.

```
make_nbinom_alist <- function(sizes, probs) {
  lapply(1:length(sizes), function(i) {
    substitute(qnbinom(size = s, prob = p),
               list(s = sizes[i], p = probs[i]))
  })
}
## make_nbinom_alist(c(20, 21, 22), c(0.3, 0.4, 0.5))
nbinom_mom <- function(x) {
  m <- mean(x)
  s <- sd(x)
  s2 <- s^2
  p <- m/s2
  r <- m^2 / (s2 - m)
  c(r, p)
}
```

Now we estimate the marginal parameters using the built-in `apply`function:

```
sizes <- apply( unname(as.matrix(brca)), 2, nbinom_mom )[1, ]
probs <- apply( unname(as.matrix(brca)), 2, nbinom_mom )[2, ]
```

Notably, the marginal NB probabilities $\hat{p}_i's$ are small — ranging in $[3.9342 \times 10^{-6}, 0.0122]$. This gives rise to highly variable counts and, typically, less restriction on potential pairwise correlation pairs. Once the functions are defined/executed to complete marginal estimation, we specify targets and generate the desired random vectors using `rvec`:

```
## Set the number of random vectors
n <- 10000
## construct margins
```

```
nb_margins <- make_nbinom_alist(sizes, probs)

## run sims

sim_nbinom <- rvec(n, nb_Rho, nb_margins, type = corType,

                   ensure_PSD = TRUE, cores = cores)

colnames(sim_nbinom) <- names(brca)
```

Figure 6 displays the simulated counts and pairwise relationships for our example genes in Table 1. Simulated counts roughly mimic the observed data but with a smoother appearance due to the assumed parameter form and with less extreme points then the observed data (c.f. Figure 1.)

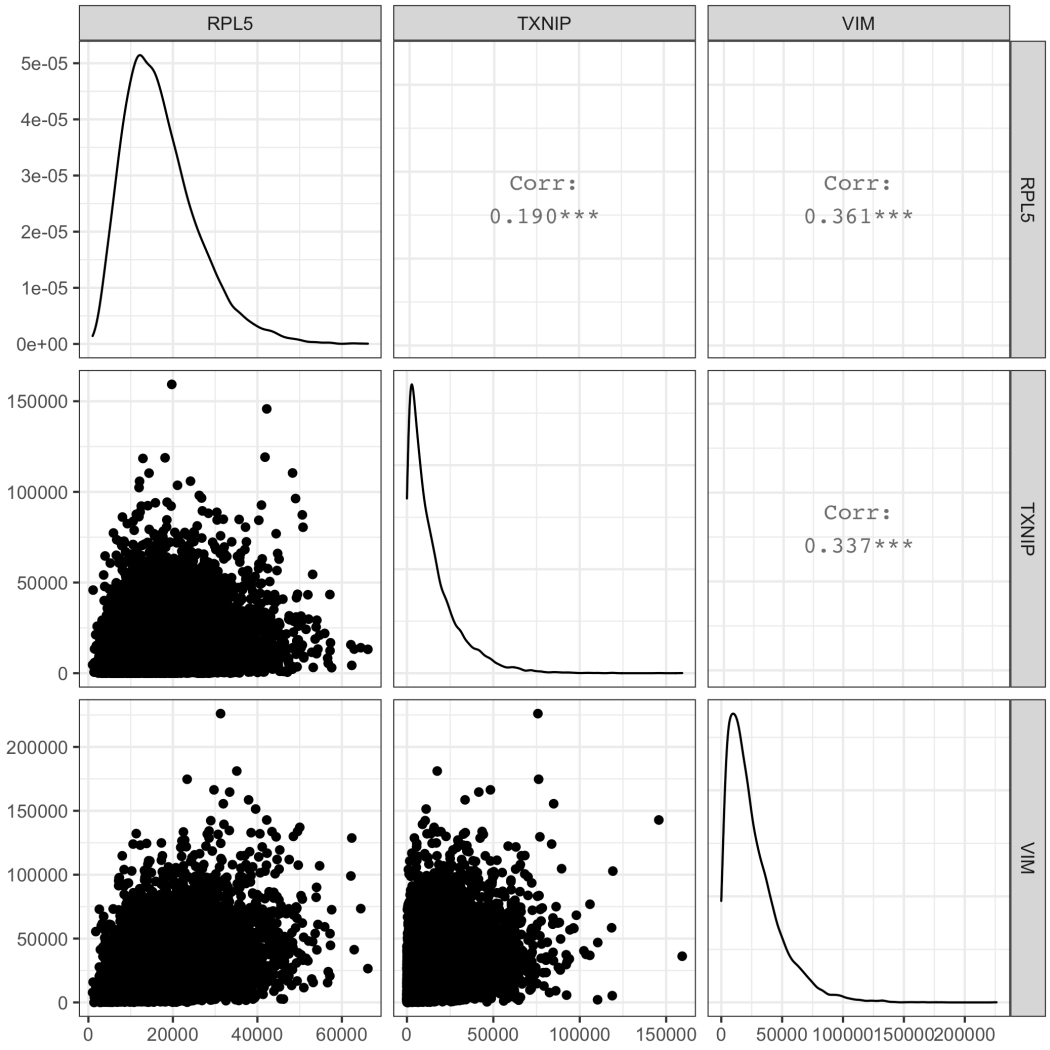

Figure 6: Simulated data for three selected high-expressing genes replicates the estimated data structure.

Figure 7 displays the aggregated results of our simulation by comparing the specified target parameter (horizontal axes) with the corresponding quantities estimated from the simulated data (vertical axes). The

evaluation shows that the simulated counts approximately match the target parameters and exhibit the full range of estimated correlation from the data. Utilizing 15 CPU threads in a MacBook Pro carrying a 2.4 GHz 8-Core Intel Core i9 processor, the simulation completed just over of 2 minutes.

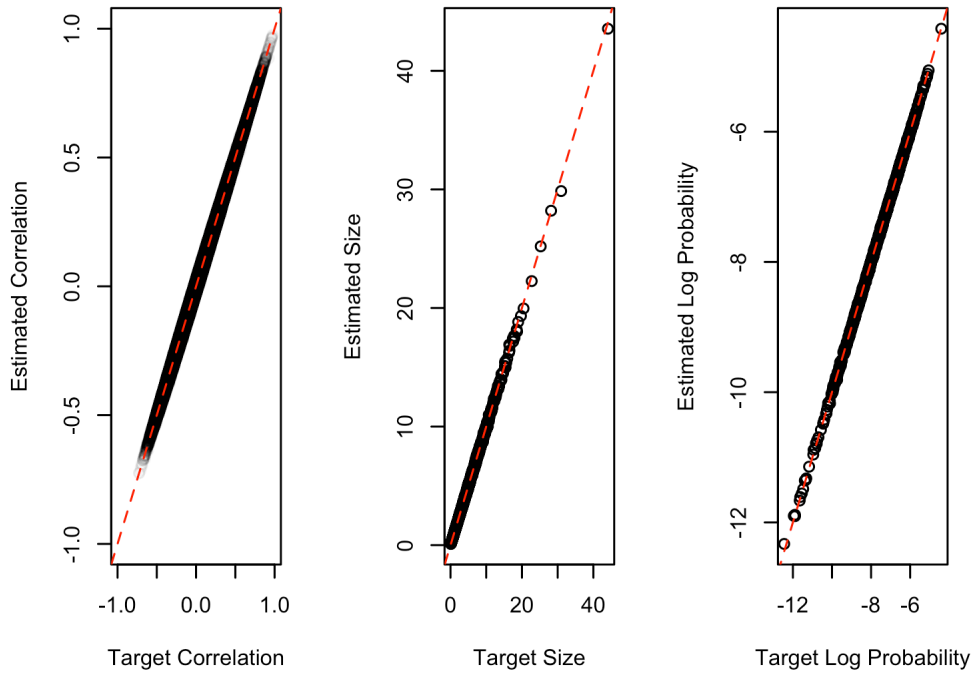

Figure 7: Simulated random vectors from a multivariate negative binomial replicate the estimated structure from an RNA-seq data set. The dashed red lines indicated equality between estimated parameters from simulated data (vertical axes) and the specified target parameters (horizontal axes).

6.2 Simulation-based joint probability calculations

To conduct statistical inference a critical task is to evaluate the joint probability mass (or density) function:

$$P(\mathbf{Y} = \mathbf{y}), y_i \in \chi_i.$$

where χ_i is the sample space for the i^{th} component of the random vector $\mathbf{Y}$. Compact representations with convenient computational forms are rare for high-dimensional constructions, especially with heterogeneous, correlated marginal distributions (or margins of mixed data types). Given a large number simulated vectors as produced above, estimated probabilities are readily given by counting the proportion of simulated vectors meeting the desired condition. In our motivating application, one may ask what is the probability that all

genes expressed greater than a certain threshold value $\mathbf{y}_0$.

Then we estimate

$$\hat{P}(\mathbf{Y} >= \mathbf{y_0}) = \sum_{b=1}^{B} I(\mathbf{Y^{(b)}} > \mathbf{y_0})/B$$

where $\mathbf{Y^{(b)}}$ is the b^{th} simulated vector in a total of B simulation replicates and $I()$ is the indicator function. For example, we can estimate from our $B = 10,000$ simulated vectors the probability that all genes are expressed (i.e., $\mathbf{y}_i \geq 1, \forall\ i$) is 0.1708.

```
d <- ncol(sim_nbinom)
B <- nrow(sim_nbinom)
threshold <- rep( 1, d)
mean(apply( sim_nbinom,  1,
           function(X, y0=threshold) {
               all( X > y0) }
           ))
```

```
[1] 0.1708
```

6.3 Evaluation of correlation estimation efficiency

MC methods are routinely used in many statistical inferential tasks including estimation, hypothesis testing, error rates, and empirical interval coverage rates. For an concise introduction to these methods, see Rizzo [2007], Ch. 6. To conclude the example applications, we demonstrate how `bigsimr` can be used evaluate estimation efficiency. In particular, we'd like to assess the error in our correlation estimation above. We used a conventional method, based on classical statistical theory. Yet this method was not designed for high-dimensional data. Indeed, high-dimensional covariance estimation (and precision matrices) is an active area of statistical science (see, for example, [Won et al., 2013, Van Wieringen and Peeters, 2016].

In this small example, we simulate $m = 10$ data sets with the number of simulated vectors matching the number of patients in the BRCA data set, $N = 1212$. Since our simulation is much faster for the Pearson correlation type (see Figure 5), we only convert the Spearman correlation matrix once (and ensure its PSD). At each iteration, we estimate the quadratic loss from the specified $\mathbf{R}_{Spearman}$, producing a distribution of loss values.

```
## Simulate random vectors equal to the sample size
n <- nrow(brca)
## convert outside for faster simulation
nb_Rho_p <- bigsimr::cor_convert( rho = nb_Rho,
                                  from = corType, to = "pearson" )
## ensure PSD
nb_Rho_p <- bigsimr::cor_nearPSD( G = nb_Rho_p )
## create m random vectors and estimate correlation
simRho <- replicate(n = m,
  expr = { tmpSim <- rvec(n = n , nb_Rho_p, nb_margins,
    type = 'pearson', ensure_PSD = FALSE, cores = cores);
    bigsimr::cor_fast( x = tmpSim, method = corType )} ,
simplify = FALSE)
## find quadratic loss at each rep
quadLoss <- unlist( lapply( simRho, rags2ridges::loss,
                            T = nb_Rho, type = "quadratic"))
```

The R summary function supplies the mean-augmented five-number summary of the quadratic loss distribution computed above.

```
summary(quadLoss)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1225076 1251629 1277417 1321153 1390574 1461040
```

This distribution could be compared to high-dimensional designed covariance estimators to guide to help decide whether the additional complexity and computation time are warranted.

7 Conclusion and discussion

We've introduced a general-purpose high-dimensional multivariate simulation algorithm and provide a high-performance R package called `bigsimr`. The random vector generation method is largely inspired by NORTA (Cario and Nelson [1997]) and Gaussian copula-based approaches (Madsen and Birkes [2013], Barbiero and Ferrari [2017], Xiao [2017]). The major contributions of this work are high-dimensional scaleability, flexible modeling of dependency, and an high-performance implementation with broad potential data analytic

applications for modern, big-data statistical computing.

It is customary to compare new tools and algorithms directly to existing competing methods and software. In this study, however, we only employ our proposed methodology, since our previous work has show that existing tools are simply not designed or feasible to meet our high-dimensional goal (see Li et al. [2019] for evaluations of the `R copula`package and others). For the bivariate simulations, existing packages such as `nortaRA` work well to match Pearson correlations exactly.

There are limitations to the methodology and implementation. The most obvious missing feature of the proposed methodology is the inability to match a Pearson correlation matrix exactly. As discussed in Algorithms and extensively by Xiao and Zhou [2019], this is a computationally intense procedure and Pearson's correlation is not a natural choice to describe dependency for non-normal marginals. While we do not provide an implementation directly supporting Pearson matching, users may supply their own input Pearson correlation after using a supplementary matching scheme [Cario and Nelson, 1997, Xiao and Zhou, 2019].

Future work includes developing scaleable algorithms to match the Pearson correlation matrix more precisely, discrete-margin specific modifications including fast Spearman's correlation rescaling (see Equation (5)), and high-dimensional covariance estimation. From an implementation standpoint, `bigsimr` only supports Nvidia GPUs and redesigning the code using OpenCL would broaden the users who would benefit. As data-analytic problems grow to even larger dimension, multi-GPU support is a promising hardware-based future direction.

Supplementary Materials

We provide an open-source implementation of our methodology as the `bigsimr` R package, hosted on github, https://schisslergroup.github.io/bigsimr/.

Acknowledgment(s)

The authors gratefully acknowledge the helpful discussions with University of Arizona's Professor Walter W. Piegorsch and Professor Edward J. Bedrick during this project's conception. We also gratefully acknowledge Heather Knudson's graphic design for the `bigsimr` R Package. The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Disclosure statement

The authors report no conflict of interest.

Funding

Research reported in this publication was supported by MW-CTR-IN of the National Institutes of Health under award number NIH 1U54GM104944.

References

Alessandro Barbiero and Pier Alda Ferrari. An R package for the simulation of correlated discrete variables. *Communications in Statistics - Simulation and Computation*, 46(7):5123–5140, aug 2017. ISSN 0361-0918. doi: 10.1080/03610918.2016.1146758. URL https://www.tandfonline.com/doi/full/10.1080/03610918.2016.1146758.

Marne C. Cario and Barry L. Nelson. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, 1997.

Huifen Chen. Initialization for NORTA: Generation of Random Vectors with Specified Marginals and Correlations. *INFORMS Journal on Computing*, 13(4):312–331, 2001.

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis, 2016. ISSN 1474760X.

Hakan Demirtas and Donald Hedeker. A practical way for computing approximate lower and upper correlation bounds. *American Statistician*, 65(2):104–109, 2011. ISSN 00031305. doi: 10.1198/tast.2011.10090.

Bradley Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103, 2007. ISSN 01621459. doi: 10.1198/016214506000001211.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*. Cambridge University Press, Cambridge, 1 edition, 2012.

Matteo Fasiolo. *An introduction to mvnfast. R package version 0.1.6.*, 2016. URL https://cran.r-project.org/package=mvnfast.

Soumyadip Ghosh and Shane G Henderson. Properties Of The Norta Method In Higher Dimensions. *Proceedings of the 2002 Winter Simulation Conference*, pages 263–269, 2002.

William H. Kruskal. Ordinal Measures of Association. *Journal of the American Statistical Association*, 53 (284):814–861, 1958. ISSN 1537274X. doi: 10.1080/01621459.1958.10501481.

Bo Li and Colin N. Dewey. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 2011. ISSN 14712105. doi: 10.1186/1471-2105-12-323.

Xiang Li, A. Grant Schissler, Rui Wu, Lee Barford, Jr. Harris, Fredrick C., and Frederick C. Harris. A Graphical Processing Unit Accelerated NORmal to Anything Algorithm for High Dimensional Multivariate Simulation. *Advances in Intelligent Systems and Computing*, pages 339–345, 2019. doi: 10.1007/978-3-030-14070-0_46.

L. Madsen and D. Birkes. Simulating dependent discrete data. *Journal of Statistical Computation and Simulation*, 2013. ISSN 00949655. doi: 10.1080/00949655.2011.632774.

Dominique Drouet Mari and Samuel Kotz. *Correlation and dependence*. World Scientific, 2001. ISBN 1860942644.

Roger B. Nelsen. *An Introduction to copulas*. Springer Science & Business Media, New York, 2 edition, 2007. ISBN 9781475719062.

Aristidis K. Nikoloulopoulos. Copula-based models for multivariate discrete response data. In *Lecture Notes in Statistics: Copulae in Mathematical and Quantitative Finance*, pages 231–249. Springer, Heidelberg, 213 edition, 2013.

Chul Gyu Park, Taesung Park, and Dong Wan Shin. A Simple Method for Generating Correlated Binary Variates. *American Statistician*, 1996. ISSN 15372731. doi: 10.1080/00031305.1996.10473557.

Houduo Qi and Defeng Sun. Computing the A Quadratically Convergent Newton Method For Computing The Nearest Correlation Matrix. *SIAM Journal on matrix analysis and applications*, 28(2):360–385, 2006.

Maria L. Rizzo. *Statistical Computing with R*. 2007. doi: 10.1201/9781420010718.

A Grant Schissler, Walter W Piegorsch, and Yves A Lussier. Testing for differentially expressed genetic pathways with single-subject N-of-1 data in the presence of inter-gene correlation. *Statistical Methods in Medical Research*, 27(12):3797–3813, 2018. ISSN 0962-2802. doi: 10.1177/0962280217712271. URL http://journals.sagepub.com/doi/10.1177/0962280217712271.

Alfred Grant Schissler, Dillon Aberasturi, Colleen Kenost, and Yves A. Lussier. A Single-Subject Method to

Detect Pathways Enriched With Alternatively Spliced Genes. *Frontiers in Genetics*, 10(414), may 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00414. URL https://www.frontiersin.org/article/10.3389/fgene.2019.00414/full.

A Sklar. Fonctions de R{é}partition {à} n Dimensions et Leurs Marges. *Publications de L'Institut de Statistique de L'Universit{é} de Paris*, 1959.

Peter Xue-kun Song. Multivariate Dispersion Models Generated from Gaussian Copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

Shing Ted Li and Joseph L. Hammond. Generation Of Pseudorandom Numbers With Specified Univariate Distributions And Correlation Coefficients. *IEEE Transactions on Systems, Man and Cybernetics*, 1975. ISSN 21682909. doi: 10.1109/TSMC.1975.5408380.

Manuel Úbeda-Flores and Juan Fernández-Sánchez. Sklar's theorem: The cornerstone of the Theory of Copulas. In *Copulas and Dependence Models with Applications*. 2017. doi: 10.1007/978-3-319-64221-5_15.

Wessel N. Van Wieringen and Carel F.W. Peeters. Ridge estimation of inverse covariance matrices from high-dimensional data. *Computational Statistics and Data Analysis*, 2016. ISSN 01679473. doi: 10.1016/j.csda.2016.05.012.

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: A revolutionary tool for transcriptomics, 2009. ISSN 14710056.

Joong-Ho Won, Johan Lim, Seung-Jean Kim, and Bala Rajaratnam. Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):427–450, jun 2013. ISSN 13697412. doi: 10.1111/j.1467-9868.2012.01049.x. URL http://doi.wiley.com/10.1111/j.1467-9868.2012.01049.x.

Di Wu and Gordon K. Smyth. Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133–e133, sep 2012. ISSN 03051048. doi: 10.1093/nar/gks461. URL https://academic.oup.com/nar/article/40/17/e133/2411151.

Qing Xiao. Generating correlated random vector involving discrete variables. *Communications in Statistics - Theory and Methods*, 2017. ISSN 1532415X. doi: 10.1080/03610926.2015.1024860.

Qing Xiao and Shaowu Zhou. Matching a correlation coefficient by a Gaussian copula. *Communications in Statistics - Theory and Methods*, 48(7):1728–1747, 2019. ISSN 1532415X. doi: 10.1080/03610926.2018.1439962.

Jun Yan. Enjoy the Joy of Copulas : With a Package copula. *Journal Of Statistical Software*, 21(4):1–21, 2007. ISSN 15487660. URL http://www.jstatsoft.org/v21/i04.