Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. doi: 10.18637/jss.v000.i00

Simulating High-Dimensional Multivariate Data using the Bigsimr Package

Alfred G. Schissler
University of Nevada

Edward J. Bedrick
University of Arizona

Alexander D. Knudson
University of Nevada

Tomasz J. Kozubowski
University of Nevada

Tin Nguyen
University of Nevada

Anna K. Panorska
University of Nevada

Juli Petereit
University of Nevada

Walter W. Piegorsch
University of Arizona

Duc Tran
University of Nevada

Abstract

It is critical to accurately simulate data when employing Monte Carlo techniques and evaluating statistical methodology. Measurements are often correlated and high dimensional in this era of big data, such as data obtained in high-throughput biomedical experiments. Due to the computational complexity and a lack of user-friendly software available to simulate these massive multivariate constructions, researchers resort to simulation designs that posit independence or perform arbitrary data transformations. To close this gap, we developed the **Bigsimr** Julia package with R and Python interfaces. These packages empower high-dimensional random vector simulation with arbitrary marginal distributions and dependency via a Pearson, Spearman, or Kendall correlation matrix. **Bigsimr** contains high-performance features, including multi-core and graphical-processing-unit-accelerated algorithms to estimate correlation and compute the nearest correlation matrix. Monte Carlo studies quantify the accuracy and scalability of our approach, up to $d = 10,000$. We describe example workflows and apply to a high-dimensional data set — RNA-sequencing data obtained from breast cancer tumor samples.

Keywords: multivariate simulation, high-dimensional data, nonparametric correlation, Gaussian copula, RNA-sequencing data, breast cancer.

1. Introduction

Massive high-dimensional (HD) data sets are now common in many areas of scientific inquiry. As new methods are developed for data analysis, 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. Further, efficient simulation empowers statistical computing strategies, such as the parametric bootstrap (Chernick 2008) to simulate from a hypothesized null model, providing inference in analytically challenging settings. Such Monte Carlo (MC) techniques become difficult for HD dependent data using existing algorithms and tools. This is particularly true when simulating massive multivariate, non-normal distributions, arising in many fields of study.

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

Another central issue lies in characterizing dependence between components in the HD 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 dependence perfectly. However, simulating arbitrary random vectors that match a target Pearson correlation matrix is computationally intense (Chen 2001; Xiao 2017). On the other hand, an analyst might consider use of nonparametric correlation measures to better characterize monotone, non-linear dependence, such as Spearman's ρ and Kendall's τ. Throughout, we focus on matching these nonparametric dependence measures, as our aim lies in modeling non-normal data and these rank-based measures possess invariance properties enabling our proposed methodology. We do, however, implement Pearson matching, but several layers of approximation are required.

With all this in mind, we present a scalable, flexible multivariate simulation algorithm. The crux of the method lies in the construction of a Gaussian copula in the spirit of the NORTA procedure. As we will describe in more detail, the algorithm's design leverages useful properties of nonparametric correlation measures, namely invariance under monotone transformation and well-known closed-form relationships between dependence measures for the multivariate normal (MVN) distribution. For our method, we developed a high-performance implementation: the `Bigsimr` Julia package, with R and Python interfaces `bigsimr`.

This article proceeds by providing background information, including a description of a motivating example application: RNA-sequencing (RNA-seq) breast cancer data. Then we describe and justify our simulation methodology and related algorithms. We proceed by providing an illustrative LD `bigsimr` workflow. Next we conduct MC studies under various bivariate distributional assumptions to evaluate performance and accuracy. After the MC evaluations, we simulate random vectors motivated by our RNA-seq example, evaluate the accuracy, and

provide example statistical computing tasks, namely MC estimation of joint probabilities and evaluating HD correlation estimation efficiency. Finally, we discuss the method's utility, limitations, and future directions.

References

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

Cario MC, Nelson BL (1997). "Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix." *Technical report*.

Chen H (2001). "Initialization for NORTA: Generation of random vectors with specified marginals and correlations." *INFORMS Journal on Computing*, **13**(4), 312–331.

Chernick MR (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers*. 2nd edition. Hoboken, NJ.

Li X, Schissler AG, Wu R, Barford L, Harris, Fredrick C J, Harris FC (2019). "A Graphical Processing Unit accelerated NORmal to Anything algorithm for high dimensional multivariate simulation." *Advances in Intelligent Systems and Computing*, pp. 339–345. doi:10.1007/978-3-030-14070-0_46.

Madsen L, Birkes D (2013). "Simulating dependent discrete data." *Journal of Statistical Computation and Simulation*, **83**(4), 677–691. ISSN 00949655. doi:10.1080/00949655.2011.632774.

Nelsen RB (2007). *An Introduction to Copulas*. 2 edition. Springer Science & Business Media, New York. ISBN 9781475719062.

Park CG, Park T, Shin DW (1996). "A simple method for generating correlated binary variates." *American Statistician*, **50**(4), 306–310. ISSN 15372731. doi:10.1080/00031305.1996.10473557.

Xiao Q (2017). "Generating correlated random vector involving discrete variables." *Communications in Statistics - Theory and Methods*, **46**(4), 1594–1605. ISSN 1532415X. doi:10.1080/03610926.2015.1024860.

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

Yan J (2007). "Enjoy the joy of copulas: with a package copula." *Journal of Statistical Software*, **21**(4), 1–21. ISSN 15487660. URL http://www.jstatsoft.org/v21/i04.

Affiliation:

Alfred G. Schissler
University of Nevada
1664 N Virginia St.
Reno, NV 89557
E-mail: aschissler@unr.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd