Generative Correlation Manifolds: Generating Synthetic Data with Preserved Higher-Order Correlations

Jens E. d'Hondt
j.e.d.hondt@tue.nl
Eindhoven University of Technology
Eindhoven, the Netherlands

Abstract

The increasing need for data privacy and the demand for robust machine learning models have fueled the development of synthetic data generation techniques. However, current methods often succeed in replicating simple summary statistics but fail to preserve both the pairwise and higher-order correlation structure of the data that define the complex, multi-variable interactions inherent in real-world systems. This limitation can lead to synthetic data that is superficially realistic but fails when used for sophisticated modeling tasks. In this white paper, we introduce Generative Correlation Manifolds (GCM), a computationally efficient method for generating synthetic data. The technique uses Cholesky decomposition of a target correlation matrix to produce datasets that, by mathematical proof, preserve the entire correlation structure – from simple pairwise relationships to higher-order interactions – of a z-normalized source dataset. We argue that this method provides a new approach to synthetic data generation with potential applications in privacy-preserving data sharing, robust model training, and simulation.

1 Introduction

In an era dominated by data-driven discovery, access to high-quality data is paramount [24, 32]. Yet, this access is often restricted by critical privacy regulations (e.g., GDPR, HIPAA) and the inherent scarcity of data in many specialized domains. Synthetic data generation offers a compelling solution, promising to provide statistically representative surrogates without exposing sensitive information or requiring new data collection [3, 5].

The central challenge, however, lies in the definition of "statistically representative". Most generative methods are validated by their ability to match the summary statistics of a source dataset [13, 19]. While valuable, this is an incomplete measure of accuracy. Real-world phenomena are rarely governed by univariate statistics or even pairwise relationships; they are driven by a web of intricate, multi-variable dependencies. In this work, we focus specifically on higher-order Pearson correlations – defined as correlations between means of variable groups – which capture one important aspect of these multi-variable dependencies [1, 2, 8, 10, 18, 21, 22, 25, 28, 29, 31, 34]. For example, (a) financial markets display cascading correlation effects during crisis periods [23], (b) biological systems contain complex gene regulatory networks with multi-order interactions [4], and (c) social networks exhibit intricate relationship structures that cannot be captured by simple pairwise correlations [27]. When synthetic data fails to preserve these structures, downstream applications suffer from reduced model performance, biased statistical analyses, and compromised privacy guarantees.

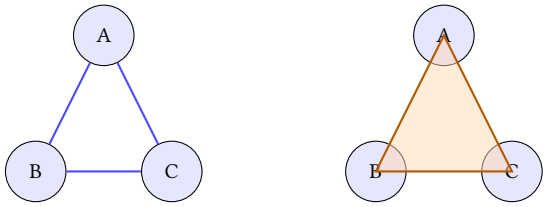

Figure 1: Illustration of the different orders of correlation.

This white paper introduces a novel approach, **Generative Correlation Manifolds (GCM)**, that directly addresses this challenge. Our method builds upon our previous work on multivariate correlation discovery [12] and extends it to synthetic data generation. Particularly, for z-normalized data, we prove that preserving pairwise correlations is mathematically sufficient to maintain all higher-order Pearson correlations. While other forms of multi-variable dependencies (e.g., mutual information or tensor-based decompositions) may exist, our method focuses specifically on this well-defined correlation structure. Based on this insight, our method leverages Cholesky decomposition to generate synthetic data that is guaranteed to preserve the complete Pearson correlation hierarchy of a z-normalized source dataset in a computationally efficient way. The implementation of GCM is available as open source software at https://github.com/JdHondt/gcm.

2 Related Work

Current synthetic data generation approaches fall into two main categories: deep learning-based methods and traditional statistical approaches. We briefly review these methods and highlight their limitations in capturing complex interactions.

Deep Learning-Based Methods. Synthetic data generation based on state-of-the-art deep learning methods has recently emerged as a promising solution to replace the expensive and laborious collection of real data. Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have shown promise in generating realistic synthetic datasets [17, 20]. However, these methods offer no mathematical guarantees regarding the preservation of relationships between features, particularly higher-order correlations. More recently, transformer-based architectures like GPT and BERT variants have been adapted for synthetic data generation [6]. While these models excel at capturing sequential patterns and contextual relationships, they similarly lack explicit guarantees for preserving correlation structures across features.

Statistical Approaches. Traditional statistical methods, including copula-based techniques, have been employed for synthetic data generation. Copulas provide a flexible framework for modeling multivariate distributions by separating the marginal distributions from their dependency structure [26]. While some copula methods can theoretically preserve the complete correlation structure through accurate modeling of pairwise dependencies, they often become computationally intractable for higher dimensions and require complex parameter estimation. Other approaches like SMOTE and ADASYN focus primarily on class balance rather than preserving feature relationships [9].

Our Contribution Our method's key contribution lies in its elegant simplicity and mathematical proof that preserving pairwise correlations is sufficient to maintain all higher-order relationships. While some existing methods may achieve similar preservation indirectly, GCM provides a direct, computationally efficient approach through its manifold-based transformation. This mathematical insight allows us to guarantee the preservation of the complete correlation structure while avoiding the complexity of explicitly modeling higher-order dependencies.

3 The Challenge of Capturing Higher-Order Correlations

To understand the importance of this work, it is crucial to distinguish between different orders of correlation.

Pairwise (2nd-Order) Correlation: This is the similarity between two variables. For example, an increase in marketing spend is correlated with an increase in sales. The most common measure of correlation is Pearson's correlation coefficient (ρ), defined for two variables x and y as:

$$\rho(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y respectively. This coefficient ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation).

Higher-Order Correlation: This involves the interaction between three or more variables. For example, consider a medical dataset where a specific gene (A), a particular lifestyle factor (B), and a negative health outcome (C) are studied. A model looking only at pairwise correlations might find weak links between A-C and B-C. However, the true risk might only become significant when both A and B are present simultaneously. This three-way interaction is a higher-order correlation. The formalization and discovery of such multivariate dependencies is a key topic in data mining research [12].

A common way to quantify these higher-order relationships is through the *Multipole* correlation measure (also known as the k-th order correlation with $k > 2$), which extends the standard correlation coefficient to multiple variables [2, 12]. The multipole correlation $MP(X)$ measures the linear dependence of an input set of vectors X [2]. Specifically, let $\hat{x}_1, \ldots, \hat{x}_n$ denote n z-normalized input (column) vectors, and $\vec{X} = [\hat{x}_1, \ldots, \hat{x}_n]$ the matrix formed by concatenating the vectors. Then:

$$MP(X) = 1 - \min_{\vec{v} \in \mathbb{R}^n, \hat{\vec{v}}=1} \text{var}(\mathbf{X} \cdot \vec{v}^T) \tag{2}$$

The value of $MP(X)$ lies between 0 and 1. The measure takes its maximum value when there exists perfect linear dependence, meaning that there exists a vector $\mathbf{v}$ with norm 1, such that $\text{var}(\mathbf{X} \cdot \vec{v}^T) = 0$.

Machine learning models, particularly deep neural networks and complex ensemble methods, can implicitly learn and exploit these higher-order structures to achieve state-of-the-art performance [33]. However, when a model is trained on synthetic data that lacks this structural richness, it learns a flawed representation of the problem space, leading to poor generalization and unreliable performance on real-world data. Therefore, synthetic data generation methods that are evaluated solely on their ability to replicate simple aggregate statistics (means, variances) may appear successful while failing to capture the complex, higher-order relationships that are critical for effective modeling. This oversight can lead to significant gaps between the performance of models trained on synthetic versus real data, particularly in domains where multi-variable interactions drive key outcomes.

4 Methodology: The GCM Method

The GCM method is elegant in its simplicity and powerful in its mathematical guarantees.

Problem Formulation Let $\hat{D} \in \mathbb{R}^{m_D \times n}$ be a z-normalized dataset of n vectors with m_D dimensions each. Define $C_{D,k} \in \mathbb{R}^{n^k}$ as the k-th order correlation matrix of $\hat{D}$. This formulation extends our previous work on multivariate correlation discovery [12], which established the theoretical foundation for identifying and measuring higher-order correlations in both static and streaming data contexts.

Objective: Generate a z-normalized synthetic dataset $\hat{S}$ with n vectors and m_S dimensions such that $C_{S,k} = C_{D,k}$ for all $k \leq m_S$.

Intuition Instead of attempting to learn a complex data distribution from scratch, our method begins with the data's relational blueprint: its *pairwise correlation matrix*. We conceptualize this matrix as defining a specific "shape" or "manifold" in a high-dimensional space. Any dataset conforming to this manifold will share the same fundamental relational properties. The GCM method uses a well-established linear algebra technique, Cholesky decomposition, as a transform. It takes unstructured, random noise and projects it onto this predefined correlation manifold. The result is a synthetic dataset that perfectly embodies the target correlation structure.

Theoretical Foundation The foundational discovery behind GCM is that for z-normalized data, all higher-order correlations are deterministic functions of the pairwise correlation matrix. This is a non-obvious but provable property. This means that if we can perfectly replicate the pairwise correlation structure, we inherently and automatically replicate the entire higher-order correlation structure for free. The method does not approximate these complex relationships; it reconstructs them exactly. Particularly, our approach is built upon the following key theorem:

Theorem 1. *Let $\hat{D} \in \mathbb{R}^{m_D \times n}$ be a z-normalized dataset with correlation matrix $C_{D,2}$. A z-normalized synthetic dataset $\hat{S}$ with n*

vectors and m_S dimensions generated through Cholesky decomposition of $C_{D,2}$ is guaranteed to have the same k-th order correlation structure as $\hat{D}$, i.e., $C_{S,k} = C_{D,k}$ for all $k \leq m_S$.

Proof. The proof relies on demonstrating that higher-order correlations can be expressed as functions of pairwise correlations for z-normalized data. This builds upon our foundational work [12] which showed that multivariate correlations in static data can be decomposed into constituent pairwise relationships. By preserving the pairwise correlation structure exactly, all higher-order structures are automatically preserved. The detailed proof is provided in Appendix A. □

Process The generation process is highly efficient and builds upon standard statistical methods for generating correlated variates [15, 30];

(1) *Extract Blueprint*: Given a source dataset, compute its $n \times n$ pairwise correlation matrix, C.
(2) *Decompose*: Perform a Cholesky decomposition on C to obtain the lower triangular matrix L, where $C = LL^T$.
(3) *Generate*: Create a matrix Z of independent random variables drawn from a standard normal distribution.
(4) *Transform*: Compute the synthetic dataset $S = ZL$.

The resulting synthetic dataset S is guaranteed to have a pairwise correlation matrix identical to C, and therefore, an identical higher-order correlation structure to the original dataset.

Computational Complexity The algorithm requires $O(n^3)$ operations for the Cholesky decomposition and $O(m_S * n^2)$ for data generation. While these complexities are non-trivial for very large datasets, the method has the advantage of being non-iterative, requiring only a single pass to generate the synthetic data once the correlation matrix is computed.

5 Use Cases and Applications

The ability to generate data with such high structural fidelity unlocks numerous possibilities:

- **Privacy-Preserving Data Sharing:** Distribute synthetic datasets that retain the full statistical utility of private source data, allowing external researchers to conduct complex modeling without ever accessing sensitive records [14].
- **Robust Model Augmentation:** Augment small or imbalanced datasets to improve the training, generalization, and fairness of machine learning models, particularly in fields like finance and medicine where feature interactions are critical [11].
- **High-Fidelity Simulation**: Create realistic, multi-variate inputs for complex systems modeling, such as financial market stress tests, epidemiological forecasting, and climate change simulations [16].
- **Algorithmic Fairness and Auditing:** Generate controlled datasets with specific correlation structures to systematically test machine learning models for bias arising from complex interactions between sensitive attributes and other features [7].

6 Call for Collaboration

The work presented here establishes the theoretical foundation of Generative Correlation Manifolds. We believe this is the first step toward a new class of synthetic data generation tools and are actively seeking collaborators to explore its potential.

We are particularly interested in pursuing research in the following areas:

- **Beyond Correlation:** Investigating the preservation of other types of higher-order correlations, such as mutual information and non-linear relationships, to further enhance the method's applicability [12].
- **Domain-Specific Applications:** Applying GCM to pressing challenges in fields like genomics, climate science, social sciences, neuroimaging, and finance, where complex multi-variable interactions are common [31]
- **Scalability and Performance:** Benchmarking the method on extremely high-dimensional datasets and optimizing its computational performance.

If you or your organization are working on challenges related to synthetic data, data privacy, or robust modeling, we invite you to connect with us.

7 Conclusion

We have presented the Generative Correlation Manifold method as a potential new approach to synthetic data generation. The method's focus on preserving the complete correlation structure of datasets offers promising opportunities for creating representative synthetic data. While further research is needed to fully understand its capabilities and limitations, initial results suggest that GCM could contribute to advancing the field of synthetic data generation, particularly in applications where preserving complex data relationships is essential.

References

[1] Saurabh Agrawal, Gowtham Atluri, Anuj Karpatne, William Haltom, Stefan Liess, Snigdhansu Chatterjee, and Vipin Kumar. [n.d.]. Tripoles: A New Class of Relationships in Time Series Data. In *Proc. SIGKDD'17*.

[2] Saurabh Agrawal, Michael Steinbach, Daniel Boley, Snigdhansu Chatterjee, Gowtham Atluri, Anh The Dang, Stefan Liess, and Vipin Kumar. 2020. Mining Novel Multivariate Relationships in Time Series Data Using Correlation Networks. *IEEE Transactions on Knowledge and Data Engineering* 32, 9 (2020), 1798–1811. https://doi.org/10.1109/TKDE.2019.2911681

[3] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. 2021. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance* (New York, New York) *(ICAIF '20)*. Association for Computing Machinery, New York, NY, USA, Article 44, 8 pages. https://doi.org/10.1145/3383455.3422554

[4] Albert-László Barabási and Zoltán N. Oltvai. 2004. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* 5, 2 (01 Feb 2004), 101–113. https://doi.org/10.1038/nrg1272

[5] André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. 2024. Comprehensive Exploration of Synthetic Data Generation: A Survey. arXiv:2401.02524 [cs.LG] https://arxiv.org/abs/2401.02524

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[7] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (01 Sep 2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[8] Örjan Carlborg and Chris S. Haley. 2004. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics* 5, 8 (01 Aug 2004), 618–625. https://doi.org/10.1038/nrg1407

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (June 2002), 321–357. https://doi.org/10.1613/jair.953

[10] Huafeng (Jason) Chen, Shaojun (Jenny) Chen, Zhuo Chen, and Feng Li. 2019. Empirical Investigation of an Equity Pairs Trading Strategy. *Manage. Sci.* 65, 1 (Jan. 2019), 370–389. https://doi.org/10.1287/mnsc.2017.2825

[11] Wuxing Chen, Kaixiang Yang, Zhiwen Yu, Yifan Shi, and C. L. Philip Chen. 2024. A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review* 57, 6 (09 May 2024), 137. https://doi.org/10.1007/s10462-024-10759-6

[12] Jens E. d'Hondt, Koen Minartz, and Odysseas Papapetrou. 2024. Efficient detection of multivariate correlations with different correlation measures. *The VLDB Journal* 33, 2 (01 Mar 2024), 481–505. https://doi.org/10.1007/s00778-023-00815-y

[13] Alvaro Figueira and Bruno Vaz. 2022. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics* 10, 15 (2022). https://doi.org/10.3390/math10152733

[14] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.* 42, 4, Article 14 (June 2010), 53 pages. https://doi.org/10.1145/1749603.1749605

[15] James E Gentle. 2003. *Random number generation and Monte Carlo methods*. Vol. 381. Springer.

[16] Paul Glasserman. 2004. *Monte Carlo methods in financial engineering*. Springer, New York. http://www.amazon.com/Financial-Engineering-Stochastic-Modelling-Probability/dp/0387004513/ref=pd_sim_b_68?ie=UTF8&refRID=1AN8JXSDGMEV2RPHFC2A

[17] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML] https://arxiv.org/abs/1406.2661

[18] Hannes Heikinheimo, Eino Hinkkanen, Heikki Mannila, Taneli Mielikäinen, and Jouni K. Seppänen. 2007. Finding low-entropy sets and trees from binary data. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Jose, California, USA) *(KDD '07)*. Association for Computing Machinery, New York, NY, USA, 350–359. https://doi.org/10.1145/1281192.1281232

[19] Elnaz Karimian Sichani, Aaron Smith, Khaled El Emam, and Lucy Mosquera. 2024. Creating High-Quality Synthetic Health Data: Framework for Model Development and Validation. *JMIR Form Res* 8 (22 Apr 2024), e53241. https://doi.org/10.2196/53241

[20] Diederik P Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML] https://arxiv.org/abs/1312.6114

[21] Arno J. Knobbe and Eric K. Y. Ho. 2006. Maximally informative k-itemsets and their efficient discovery. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA, USA) *(KDD '06)*. Association for Computing Machinery, New York, NY, USA, 237–244. https://doi.org/10.1145/1150402.1150431

[22] Silvan Licher, Shahzad Ahmad, Hata Karamujić-Čomić, Trudy Voortman, Maarten J. G. Leening, M. Arfan Ikram, and M. Kamran Ikram. 2019. Genetic predisposition, modifiable-risk-factor profile and long-term dementia risk in the general population. *Nature Medicine* 25, 9 (2019), 1364–1369.

[23] François Longin and Bruno Solnik. 2001. Extreme Correlation of International Equity Markets. *The Journal of Finance* 56, 2 (2001), 649–676. http://www.jstor.org/stable/222577

[24] Yingzhou Lu, Lulu Chen, Yuanyuan Zhang, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. 2025. Machine Learning for Synthetic Data Generation: A Review. arXiv:2302.04062 [cs.LG] https://arxiv.org/abs/2302.04062

[25] Ileena Mitra, Alinoë Lavillaureix, Erika Yeh, Michela Traglia, Kathryn Tsang, Carrie E. Bearden, Katherine A. Rauen, and Lauren A. Weiss. 2017. Reverse Pathway Genetic Approach Identifies Epistasis in Autism Spectrum Disorders. *PLOS Genetics* 13 (01 2017), 1–27.

[26] Roger B. Nelsen. 2010. *An Introduction to Copulas*. Springer Publishing Company, Incorporated.

[27] M. E. J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Rev.* 45, 2 (Jan. 2003), 167–256. https://doi.org/10.1137/s003614450342480

[28] Hoang Vu Nguyen, Emmanuel Müller, Periklis Andritsos, and Klemens Böhm. [n.d.]. Detecting Correlated Columns in Relational Databases with Mixed Data Types. In *Proc. SSDBM'14*.

[29] Marcelo Perlin. 2007. M of a kind: A Multivariate Approach at Pairs Trading. *Available at SSRN 952782* (2007).

[30] Reuven Y Rubinstein and Dirk P Kroese. 2016. *Simulation and the Monte Carlo method*. John Wiley & Sons.

[31] Andrea Santoro, Federico Battiston, Giovanni Petri, and Enrico Amico. 2023. Higher-order organization of multivariate time series. *Nature Physics* 19, 2 (01 Feb 2023), 221–229. https://doi.org/10.1038/s41567-022-01852-0

[32] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. 2023. Beyond neural scaling laws: beating power law scaling via data pruning. arXiv:2206.14486 [cs.LG] https://arxiv.org/abs/2206.14486

[33] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & Cross Network for Ad Click Predictions. In *Proceedings of the ADKDD'17* (Halifax, NS, Canada) *(ADKDD'17)*. Association for Computing Machinery, New York, NY, USA, Article 12, 7 pages. https://doi.org/10.1145/3124749.3124754

[34] Xiang Zhang, Feng Pan, Wei Wang, and Andrew Nobel. 2008. Mining non-redundant high order correlations in binary data. *Proc. VLDB Endow.* 1, 1 (Aug. 2008), 1178–1188. https://doi.org/10.14778/1453856.1453981

A Formal Proof: Generation of Data with Specific Higher-order Correlation Structure via Cholesky Decomposition

A.1 Theorem

Let $\hat{D} \in \mathbb{R}^{m_D \times n}$ be a z-normalized dataset of n vectors with m_D dimensions each, and let $C_{D,k} \in \mathbb{R}^{n^k}$ be the k-th order correlation matrix (symmetric positive semi-definite with ones on the diagonal) of $\hat{D}$. Then, a z-normalized synthetic dataset $\hat{S}$ with n vectors and m_S dimensions generated through Cholesky decomposition of $C_{D,2}$ is guaranteed to have the same k-th order correlation structure as $\hat{D}$, i.e., $C_{S,k} = C_{D,k}$ for all $k \leq m_S$.

A.2 Definitions

- Let the Multipole correlation be defined as in Equation 2.
- Let z-normalization of a vector x be defined as: $\hat{x} = \frac{x - \bar{x}}{\sigma_x}$, where $\bar{x}$ is the mean and σ_x is the standard deviation of x.
- Let $Z \in \mathbb{R}^{m \times n}$ be a matrix whose rows are independent random vectors $z_i \in \mathbb{R}^n$ with $\mathbb{E}[z_i] = 0$ and $\mathrm{Cov}(z_i) = I_n$.
- Let $X = ZL$ where L is the Cholesky factor of C.
- Let $\hat{C}_{X,k}$ be the sample correlation matrix computed from the m_X samples in X.

A.3 Lemmas

Lemma 1 (K-order Correlation as a Function of Pairwise Correlations). *The multipole correlation $MP(X)$ of a set of z-normalized vectors $X = [x_1, ..., x_k]$ can be rewritten as [2]:*

$$MP(X) = 1 - \lambda_{\min}(C_{X,2}) \tag{3}$$

where $\lambda_{\min}(C_{X,2})$ is the smallest eigenvalue of the second-order correlation matrix $C_{X,2}$.

Proof. We refer to [2] for the proof of this lemma. □

A.4 Main Proof

We need to show that the k-th order correlation of the generated dataset $\hat{S}$ is equal to the k-th order correlation of the original dataset $\hat{D}$.

A.4.1 Step 1: Generating Data with a Given Pairwise Correlation Matrix Given a target correlation matrix $C_{D,2} \in \mathbb{R}^{n \times n}$, we can generate a synthetic dataset $S \in \mathbb{R}^{m_S \times n}$ with the desired correlation structure using the Cholesky decomposition method. The procedure is as follows:

(1) Compute the Cholesky decomposition of the correlation matrix $C_{D,2} = LL^T$, where L is a lower triangular matrix.
(2) Generate a matrix $Z \in \mathbb{R}^{m_S \times n}$ of independent random variables with standard normal distribution.
(3) Compute $S = ZL$.

The resulting matrix S will have the correlation structure specified by $C_{D,2}$ as m_S approaches infinity. This is because the expected correlation matrix of S is:

$$E[S^T S] = E[(ZL)^T (ZL)] \tag{4}$$
$$= E[L^T Z^T ZL] \tag{5}$$
$$= L^T E[Z^T Z] L \tag{6}$$
$$= L^T I L \tag{7}$$
$$= L^T L \tag{8}$$
$$= C_{D,2} \tag{9}$$

This approach is well-established in the statistical literature [15, 30] and provides a direct method for generating data with a specified correlation structure.

A.4.2 Step 2: Normalizing the Generated Data While the generation method in Step 1 ensures that the resulting dataset S has the desired pairwise correlation structure in expectation, this does not mean that the higher-order correlation structure is preserved. To ensure that the generated dataset S has the same higher-order correlation structure as the original dataset $\hat{D}$, we need to apply z-normalization to each column of S.

Note that we can apply z-normalization to obtain $\hat{S}$ without affecting the correlation structure, because Pearson correlation is invariant to linear transformations of the form $ax + b$ where $a > 0$.

For each column vector s_j in S, we compute:

$$\hat{s}_j = \frac{s_j - \bar{s}_j}{\sigma_{s_j}} \tag{10}$$

Since the Pearson correlation coefficient between two vectors u and v is defined as:

$$\rho(u, v) = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2 \sum_{i=1}^{n}(v_i - \bar{v})^2}} \tag{11}$$

It is invariant to z-normalization, as z-normalization is equivalent to applying the linear transformation $a = \frac{1}{\sigma_x}$ and $b = -\frac{\bar{x}}{\sigma_x}$ to each column. Therefore, $\rho(\hat{s}_i, \hat{s}_j) = \rho(s_i, s_j)$ for all columns i and j.

Thus, the z-normalized dataset $\hat{S}$ still preserves the pairwise correlation structure of the original dataset $\hat{D}$, i.e., $C_{S,2} = C_{D,2}$.

A.4.3 Step 3: Preserving Higher-order Correlation Structure To complete our proof, we need to show that the higher-order correlation structure is also preserved, i.e., $C_{S,k} = C_{D,k}$ for all $k \leq m_S$.

By Lemma 1, the k-th order correlation between any two sets of vectors $\hat{X} = [\hat{x}_1, ..., \hat{x}_s]$ and $\hat{Y} = [\hat{y}_1, ..., \hat{y}_t]$ with $s + t = k$ can be expressed solely in terms of their pairwise correlations:

$$MP(\hat{X}) = 1 - \lambda_{\min}(C_{X,2}) \tag{12}$$

Since we have established that $\hat{S}$ has the same pairwise correlation structure as $\hat{D}$ (i.e., $\rho(\hat{s}_i, \hat{s}_j) = \rho(\hat{d}_i, \hat{d}_j)$ for all columns i and j), the k-th order correlation between any two sets of vectors will also be identical.

For any combination of k vectors from $\hat{S}$, the k-th order correlation will be computed using the same pairwise correlations as the corresponding vectors in $\hat{D}$. Therefore, by the formula in Lemma 1, we can conclude that the k-th order correlation structures are identical, i.e., $C_{S,k} = C_{D,k}$ for all $k \leq m_S$.