mathematics

MDPI

*Article*

# A Generative Model for Correlated Graph Signals

**Pavel Loskot**

ZJU-UIUC Institute, Haining 314400, China; pavelloskot@intl.zju.edu.cn; Tel.: +86-571-8757-2579

**Abstract:** A graph signal is a random vector with a partially known statistical description. The observations are usually sufficient to determine marginal distributions of graph node variables and their pairwise correlations representing the graph edges. However, the curse of dimensionality often prevents estimating a full joint distribution of all variables from the available observations. This paper introduces a computationally effective generative model to sample from arbitrary but known marginal distributions with defined pairwise correlations. Numerical experiments show that the proposed generative model is generally accurate for correlation coefficients with magnitudes up to about 0.3, whilst larger correlations can be obtained at the cost of distribution approximation accuracy. The generative models of graph signals can also be used to sample multivariate distributions for which closed-form mathematical expressions are not known or are too complex.

**Keywords:** covariance; generative model; graph signal; probability distribution; random vector



**Citation:** Loskot, P. A Generative Model for Correlated Graph Signals. *Mathematics* **2021**, *9*, 3078. https://doi.org/10.3390/math9233078

Academic Editors: Irina Cristea and Frank Werner

Received: 31 October 2021
Accepted: 27 November 2021
Published: 29 November 2021



## 1. Introduction

The observations of many real-world systems can be studied as multiple time series. Provided that the pairwise relationships between the time series are implicitly or explicitly defined, it is common to refer to these data models as graph signals [1]. In most cases, only one feature representing the pairwise relationships is considered. The pairwise relationships, which are not explicitly stated can assume some implicit value, such as having a zero covariance (i.e., being uncorrelated), or these relationships should be assumed to be undefined (i.e., unknown). The graph edges can also indicate statistical or causal dependencies [2]. Consequently, graph signals can be defined as random vectors with incomplete knowledge of their statistics. The random variables then represent nodes of the graph, and the pairwise relationships are the graph edges.

The graph variables can be arranged into a random vector or matrix in an arbitrary order. One can also assume a graph search and follow a path over the graph edges through the graph nodes to construct the random vector. The random vectors and matrices can be conveniently processed using the well-established framework of linear algebra combined with methods in statistical signal processing [3] and machine learning. In the literature on graph signal processing, the mainstream approach assumes a frequency domain representation of graph signals using the singular value decomposition (SVD) of graph adjacency, incidence, or Laplacian matrices [4].

In general, a random vector or matrix is fully statistically described by a joint distribution of all the constituent random variables. A comprehensive survey of multivariate distributions can be found in [5]. Mathematical expressions of multivariate distributions may contain the pairwise correlations explicitly, as is the case of multivariate normal distribution, but in most cases, the correlations can be calculated from other distribution parameters. The main challenge in fitting multivariate distributions to observed data is that the number of required observations to achieve a certain goodness-of-fit grows exponentially with the number of dimensions (the curse of dimensionality). In addition, multivariate distributions are often difficult to sample, and numerically complex algorithms are required to perform statistical inferences [6].

A related problem of generating random variables with defined moments was considered in the literature. In particular, a linear transformation was utilized in [7,8] to generate nonnormal variates with the defined univariate skewness, kurtosis, and pairwise covariances. This method was extended in [9] to assume multivariate skewness and kurtosis.

In this paper, it is assumed that a random vector is described by the marginal distributions and the pairwise correlations of its elements. The task is to define a generative model to efficiently generate multivariate samples satisfying the given statistical constraints without resorting to fitting a multivariate distribution to the observed data. It is proposed to approximate the unknown multivariate distribution with defined statistical constraints by a multivariate mixture distribution having independent marginals. In addition, it is shown that the component distributions of the marginal mixture distributions can be conjugate distributions. This approach enables a definition of a universal procedure for constructing generative models of multivariate graph signals that can be readily sampled. The limitation of the proposed procedure is that there is a tradeoff in how accurately the marginal distributions can be approximated and the achievable pairwise correlations. However, it is likely that the proposed generative procedure could be further modified to improve the tradeoff.

The rest of the paper is organized as follows. The research problem is stated in Section 2. The generative model of graph signals is introduced in Section 3. Numerical examples for bivariate distributions are presented in Section 4. The obtained results are discussed in Section 5, and Section 6 concludes the paper.

## 2. Problem Statement

Assume that a sufficient number of discrete-time observations of $N$ stochastic time series, $X_i$, $i \in \{1, 2, \ldots, N\}$, have been collected, so the following quantities can be determined with a good accuracy:

$$\text{marginal densities, } f_i(X) \text{ of } X_i, \text{ for } \forall i \in \{1, 2, \ldots, N\}; \text{ and} \quad \text{(C1)}$$

$$\text{covariances, } C_{ij} = \operatorname{cov}[X_i, X_j] = \mathrm{E}\big[(X_i - \bar{X}_i)(X_j - \bar{X}_j)\big], \text{ for } \forall i, j \in \{1, 2, \ldots, N\} \quad \text{(C2)}$$

where $X_i$ is a random variable representing the samples in the $i$-th time series, and $\mathrm{E}[\cdot]$ denotes expectation. Note that the existence of time-invariant densities implies stationarity as well as knowledge of all the moments of univariate random variable $X_i$. Furthermore, only continuous multivariate distributions are considered in this paper, and, unless otherwise stated, $X_i \in \mathcal{R}$.

If the constraints C1 and C2 are determined from real-world observations with sufficient accuracy, both constraints are guaranteed to be consistent, and there must exist at least one corresponding multivariate distribution. However, given a set of marginal distributions and (unnormalized) covariances, there may be, in general, no multivariate distributions satisfying these constraints. The constraint C2 may be modified to assume the normalized correlation coefficients instead of covariances.

The joint moments including covariances and the corresponding Pearson correlation coefficients can be estimated from data by the method of sample moments [10]. The marginal densities can be efficiently estimated from data by various nonparametric methods using histograms, kernels [11], and diffusion [12]. However, estimating the joint probability density of all $N$ variates $X_i$ may be problematic, since it may require a very large number of observations, especially for more complex distributions in many dimensions [13]. Consequently, given the marginal distributions $f_i(X)$ and the pairwise covariances $C_{ij}$, the task is to construct a generative model for generating random samples of the vector, $\boldsymbol{X} = [X_1, X_2, \ldots, X_N]$, whose elements satisfy the constraints C1 and C2 given above. No other constraints are adopted in this paper, although it may be desirable to require that the generative model is also numerically efficient. In addition, the generative procedure to obtain random samples satisfying the constraints C1 and C2 should be sufficiently general in order to allow for different types of marginal distributions including the case of non-identical marginal distributions.

## 3. Constructing a Multivariate Distribution from Its Marginals

The constraints C1 and C2 do not uniquely define the joint distribution $f(\boldsymbol{X})$. In particular, the marginal distributions can be used to obtain all general and central moments of individual random variables $X_i$, whereas the pairwise covariances are the only joint statistics assumed to be known. Provided that the marginals $f_i$ are of the same and more common type, the corresponding multivariate density may have been identified in the literature [5]. However, even if the mathematical expression of the desired multivariate distribution $f(\boldsymbol{X})$ is available, it may be too complex to sample from, or to accurately fit to the observations using, for example, the least squares regression or other parameter estimation methods [3]. Another strategy, which is investigated in this paper, is to construct the joint distribution $f$ from the known marginals $f_i$, $i = 1, 2, \ldots, N$ under the covariance constraints.

**Proposition 1.** *The join density $f$ with the given marginals $f_i$, $i = 1, 2, \ldots, N$, under mild covariance constraints can be well approximated by the mixture distribution,*

$$f(\boldsymbol{X}) = \sum_{k=1}^{K} \alpha_k \tilde{f}_k(\boldsymbol{X}) \tag{1}$$

*of K joint component densities $\tilde{f}_k$, and the weighting factors, $\alpha_k > 0\ \forall k$ and $\sum_{k=1}^{K} \alpha_k = 1$.*

The mixture decomposition (1) of $f$ again requires that not only all the components $\tilde{f}_k$ are identified but also that they can be effectively sampled from. In order to overcome the latter challenge, it is newly proposed to adopt an independence assumption, and express the marginals $\tilde{f}_k$ as a product of the individual densities, i.e., the mixture decomposition (1) is rewritten as,

$$f(\boldsymbol{X}) = \sum_{k=1}^{K} \alpha_k \prod_{i=1}^{N} \tilde{f}_{ki}(X_i). \tag{2}$$

The advantage of assuming the mixture decomposition (2) is that it is generally much easier to sample from univariate than from multivariate distributions. The disadvantage of decomposition (2) is that the independence assumption limits the achievable pairwise correlations between variables $\boldsymbol{X}$.

Denote $\boldsymbol{X}_{-i} = \{X_1, X_2, \ldots, X_N\} \setminus \{X_i\}$, and $\boldsymbol{X}_{-i-j} = \{X_1, X_2, \ldots, X_N\} \setminus \{X_i, X_j\}$. The marginal distributions corresponding to decomposition (2) are obtained as

$$f_i(X_i) = \int_{\mathcal{R}^{N-1}} f(\boldsymbol{X})\, \mathrm{d}\boldsymbol{X}_{-i} = \sum_{k=1}^{K} \alpha_k \tilde{f}_{ki}(X_i), \tag{3}$$

whilst the bivariate marginal densities are computed as

$$f_{ij}(X_i, X_j) = \int_{\mathcal{R}^{N-2}} f(\boldsymbol{X})\, \mathrm{d}\boldsymbol{X}_{-i-j} = \sum_{k=1}^{K} \alpha_k \tilde{f}_{ki}(X_i) \tilde{f}_{kj}(X_j). \tag{4}$$

The corresponding mean value is

$$\bar{X}_i = \mathrm{E}[X_i] = \sum_{k=1}^{K} \alpha_k \mathrm{E}_{\tilde{f}_{ki}}[X_i] = \sum_{k=1}^{K} \alpha_k \bar{X}_{ki}, \tag{5}$$

and the second moments are computed as

$$\begin{aligned}
\mathrm{var}[X_i] &= \mathrm{E}[X_i^2] - \mathrm{E}[X_i]^2 = \sum_{k=1}^{K} \mathrm{E}_{\tilde{f}_{ki}}[X_i^2] - \sum_{k=1}^{K} \sum_{l=1}^{K} \alpha_k \alpha_l \bar{X}_{ki} \bar{X}_{li} \\
\mathrm{corr}[X_i, X_j] &= \mathrm{E}[X_i X_j] = \sum_{k=1}^{K} \alpha_k \bar{X}_{ki} \bar{X}_{kj} \\
\mathrm{cov}[X_i, X_j] &= \mathrm{E}[X_i X_j] - \mathrm{E}[X_i]\mathrm{E}[X_j] = \sum_{k=1}^{K} \alpha_k \bar{X}_{ki} \bar{X}_{kj} - \sum_{k=1}^{K} \sum_{l=1}^{K} \alpha_k \alpha_l \bar{X}_{ki} \bar{X}_{lj}.
\end{aligned} \tag{6}$$

Consequently and importantly, for $N \geq 2$ and $K \geq 2$, the pairwise covariances are given by the mixture coefficients $\alpha_k$ and the mean values $\bar{X}_{ki}$ of the component distributions $\tilde{f}_{ki}(X_i)$.

Given the marginals $f_i$ and the covariances $\mathrm{cov}[X_i, X_j]$, the values of $\alpha_k$ and $\bar{X}_{ki}$ must satisfy Equations (5) and (6). This represents a system of $(N + \binom{N}{2})$ equations with $(NK + K)$ unknowns, where $\binom{N}{2} = N(N-1)/2$ is a binomial coefficient. Provided that the $K$ coefficients $\alpha_k$ can be determined by $N$ univariate decompositions (3), the number of unknowns can be reduced to $NK$. Then the number of degrees of freedom as a difference between the number of equations and the number of unknowns, i.e., $D_{\text{free}} = NK - N - \binom{N}{2}$, versus the distribution dimension $N$ is shown in Figure 1 for several different values of $K$. The points above the horizontal dashed line in Figure 1 indicate the underdetermined cases when the number of unknowns is greater than the number of constraints. The unique solution may exist, if $N = 2(K-1) + 1$, i.e., when $N$ is odd.

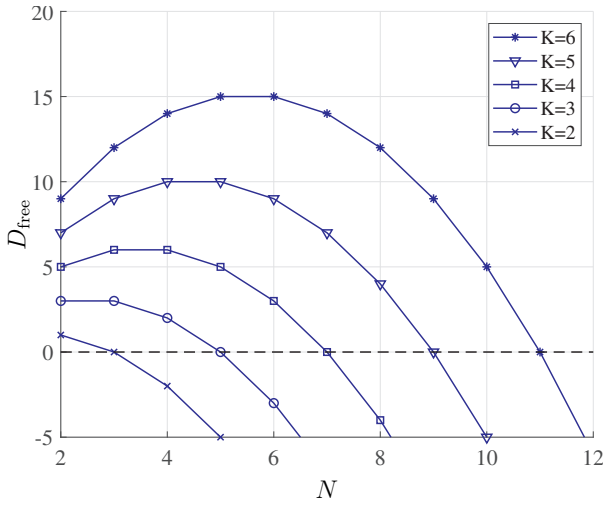


**Figure 1.** The number of degrees of freedom to determine $NK$ unknown coefficients $\bar{X}_{ki}$ from the constraints C1 and C2.

*Bivariate Case*

For $N = 2$ and $K = 2$, the number of degrees of freedom, $D_{\text{free}} = 1$. Assuming two random variables $X$ and $Y$, the mixture decomposition (2) can be rewritten as,

$$f(X,Y) = \alpha \tilde{f}_{X_1}(X)\tilde{f}_{Y_1}(Y) + (1-\alpha)\tilde{f}_{X_2}(X)\tilde{f}_{Y_2}(Y) \tag{7}$$

where $0 < \alpha < 1$. Note that, for $\alpha = 0$ or $\alpha = 1$, the variables $X$ and $Y$ are assumed to be independent, and thus, uncorrelated. The corresponding marginal distributions are the mixtures

$$\begin{aligned} f_X(X) &= \alpha \tilde{f}_{X_1}(X) + (1-\alpha)\tilde{f}_{X_2}(X) \\ f_Y(Y) &= \alpha \tilde{f}_{Y_1}(Y) + (1-\alpha)\tilde{f}_{Y_2}(Y) \end{aligned} \tag{8}$$

having the following first and second order statistics

$$\begin{aligned} \bar{X} &= \alpha \bar{X}_1 + (1-\alpha)\bar{X}_2 \\ \bar{Y} &= \alpha \bar{Y}_1 + (1-\alpha)\bar{Y}_2 \\ \mathrm{var}[X] &= \alpha\, \mathrm{var}[X_1] + (1-\alpha)\mathrm{var}[X_2] + \alpha(1-\alpha)(\bar{X}_1 - \bar{X}_2)^2 \\ \mathrm{var}[Y] &= \alpha\, \mathrm{var}[Y_1] + (1-\alpha)\mathrm{var}[Y_2] + \alpha(1-\alpha)(\bar{Y}_1 - \bar{Y}_2)^2 \\ \mathrm{corr}[X,Y] &= \alpha \bar{X}_1 \bar{Y}_1 + (1-\alpha)\bar{X}_2 \bar{Y}_2 \\ \mathrm{cov}[X,Y] &= \alpha(1-\alpha)(\bar{X}_1 - \bar{X}_2)(\bar{Y}_1 - \bar{Y}_2) \end{aligned} \tag{9}$$

where the mean and the variance of $\tilde{f}_{X_1}$ are denoted as $\bar{X}_1$ and $\mathrm{var}[X_1]$, respectively. Similar notation is used for the other component distributions. Note that the correlation between

the variables $X$ and $Y$ increases with the difference of the means of the corresponding component distributions. Conversely, the variables $X$ and $Y$ are uncorrelated, if either the means $\bar{X}_1 = \bar{X}_2$, or $\bar{Y}_1 = \bar{Y}_2$. The expression, $\alpha(1-\alpha) \geq 0$, is maximized for $\alpha = 1/2$.

Since $D_{\text{free}} = 1$, the mean values of the component distributions in (7) can be expressed as functions of one parameter. For example, choosing $\bar{X}_1$ as such parameter, the other means are computed as

$$\begin{aligned} \bar{X}_2 &= \frac{\bar{X} - \alpha \bar{X}_1}{1-\alpha} \\ \bar{Y}_1 &= \frac{\text{cov}[X,Y](1-\alpha) + \alpha \bar{Y}(\bar{X}_1 - \bar{X})}{\alpha(\bar{X}_1 - \bar{X})} \\ \bar{Y}_2 &= \frac{-\text{cov}[X,Y] + \bar{Y}(\bar{X}_1 - \bar{X})}{(\bar{X}_1 - \bar{X})} \end{aligned} \tag{10}$$

by solving the equations for $\bar{X}$, $\bar{Y}$ and $\text{cov}[X,Y]$ given in (9). Substituting the expressions in (9), the Pearson correlation coefficient is computed as,

$$\rho_{XY} = \frac{\text{cov}[X,Y]}{\sqrt{\text{var}[X]\text{var}[Y]}} = \frac{\alpha(1-\alpha)(\bar{X}_1 - \bar{X}_2)(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\sigma_1^2 + \alpha(1-\alpha)(\bar{X}_1 - \bar{X}_2)^2}\sqrt{\sigma_2^2 + \alpha(1-\alpha)(\bar{Y}_1 - \bar{Y}_2)^2}} \tag{11}$$

where $\sigma_1^2 = \alpha\,\text{var}[X_1] + (1-\alpha)\text{var}[X_2]$ and $\sigma_2^2 = \alpha\,\text{var}[Y_1] + (1-\alpha)\text{var}[Y_2]$. Consequently, the combined variances $\sigma_1^2$ and $\sigma_2^2$ limit the achievable correlations between the variables $X$ and $Y$ in the generative model (7). Only when $\sigma_1^2 = \sigma_2^2 = 0$, can the correlation coefficient reach the maximum magnitude, $\rho_{XY} = \pm 1$.

The mixture decompositions of marginals defined in (8) can be obtained using different strategies. The marginal distributions defined by the constraint C1 can be common univariate distributions, or they can be defined as the univariate mixture distributions from the onset. The latter case can be resolved by curve fitting to the observed data, so here, we investigate the former case.

**Proposition 2.** *The marginal distributions (8) can be approximated by conjugate mixtures. The conjugate mixtures are of the same type as the resulting marginal distributions, but they have their parameters determined by the constraints (9).*

Hence, assume that the mixture distributions $\tilde{f}_{X_1}$ and $\tilde{f}_{X_2}$ in (8) are obtained by a linear transformation of the marginal distribution $f_X$, i.e., let [14]

$$\begin{aligned} \tilde{f}_{X_1}(X) &= \frac{1}{s_1} f_X\left(\frac{X - m_1}{s_1}\right) \\ \tilde{f}_{X_2}(X) &= \frac{1}{s_2} f_X\left(\frac{X - m_2}{s_2}\right) \end{aligned} \tag{12}$$

where the shifts $m_1, m_2 \in \mathcal{R}$, whereas the scaling $s_1 = 1 - \epsilon_1$ and $s_2 = 1 - \epsilon_2$, for some small $\epsilon_1, \epsilon_2 > 0$, to satisfy the variance constraint in (9). The marginal distribution $f_Y$ is approximated similarly, and independently from $f_X$.

Substituting (12) into (8), the value of the mixture coefficient $\alpha$ can be determined to optimize the goodness of fit. In particular, the conjugate mixture distributions can be locally linearized about $X_0$, as indicated in Figure 2. Then, for $\forall X : |X - X_0| < \epsilon$, the distributions can be approximated as linear functions, i.e.,

$$\begin{aligned} f_X(X) &\approx gX + o \\ f_{X_1}(X) &\approx g_1(X + m_1) + o \\ f_{X_2}(X) &\approx g_2(X - m_2) + o \end{aligned} \tag{13}$$

where $g$, $g_1$, and $g_2$ are the gradients, $o$ is the common offset, and $m_1, m_2 > 0$ are the shifts. Substituting into (8), we obtain

$$\begin{aligned} \alpha\, g_1(1 + m_1/m_2) &= g \\ (1-\alpha) g_2(1 + m_2/m_1) &= g, \end{aligned} \tag{14}$$

which is crucially independent of the actual value of $X_0$. In the case when $g_1 = g_2$, a rule of thumb for choosing the value of mixture coefficient $\alpha$ is obtained as

$$m_1 \alpha = m_2(1-\alpha) \tag{15}$$

so that, $\alpha(\bar{X} + m_1) + (1-\alpha)(\bar{X} - m_2) = \bar{X}$ as in (9). Hence, the value of $\alpha$ can be chosen somewhat arbitrarily as long as the condition (15) and the constraints (9) are satisfied.

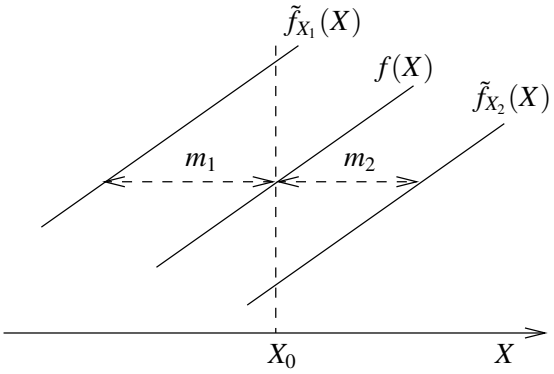


**Figure 2.** Linearization of distributions in the mixture decomposition of the marginal distribution $f_X$ in the vicinity of an arbitrary point $X_0$.

Alternatively, for conjugate mixture components, the decomposition (8) can be rewritten as,

$$f_X(X) \approx \alpha g_1 f_X(X + \Delta X) + (1-\alpha) g_2 f_X(X - \Delta X). \tag{16}$$

Equation (16) can be assumed to be a linear digital filtering of the signal $f_X(X)$ in variable $X$. The filter coefficients $\alpha g_1$ and $(1-\alpha)g_2$ are separated by $2\Delta X$. The approximation (16) is then more exact, provided that the filter does not distort the filtered signal $f_X(X)$, i.e., when the filter bandwidth is wider than the bandwidth of the signal [15]. The signal and filter bandwidth are determined by the magnitude of the Fourier transform. In particular, since the signal $f_X(X)$ is also a distribution, we can assume the characteristic function of $f_X(X)$, i.e., $\phi_X(s) = \mathrm{E}_X\left[\mathrm{e}^{\mathrm{j}sX}\right]$, $\mathrm{j} = \sqrt{-1}$, which is known or can be obtained for most univariate distributions. The filter bandwidth is obtained by computing the magnitude of its transfer function $T(s)$, i.e.,

$$\begin{aligned} T(s) &= \int_{-\infty}^{\infty} \left\{ \alpha g_1 \delta(X + \Delta X) + (1-\alpha) g_2 \delta(X - \Delta X) \right\} \mathrm{e}^{\mathrm{j}sX}\, \mathrm{d}X \\ &= \alpha g_1\, \mathrm{e}^{\mathrm{j}2\pi s \Delta X} + (1-\alpha) g_2\, \mathrm{e}^{\mathrm{j}2\pi s \Delta X}. \end{aligned} \tag{17}$$

## 4. Numerical Examples

The case of the following three bivariate distributions is considered: normal, gamma, and normal-exponential distributions [14]. Although generating correlated normal samples is straightforward, which is a rare exception among multivariate distributions, the normal distribution is mainly considered to validate the proposed generative model.

The first experiment investigates approximations of the selected univariate distributions by a mixture of the two component distributions defined in (8), i.e., the approximation, $f_{\tilde{X}} = \alpha \tilde{f}_{X_1} + (1-\alpha)\tilde{f}_{X_2}$, of $f_X$. The approximation accuracy is quantified by the Kullback–

Leibler (KL) divergence between the target distribution $f_X$ and its mixture approximation $f_{\tilde{X}}$. The KL divergence is defined as,

$$\mathrm{KL}(f_{\tilde{X}} \| f_X) = \int_{-\infty}^{\infty} f_{\tilde{X}}(X) \log \frac{f_{\tilde{X}}(X)}{f_X(X)} \, \mathrm{d}X. \tag{18}$$

Moreover, using Jensen's inequality for logarithm, it is straightforward to show that

$$\mathrm{KL}(f_X \| f_{\tilde{X}}) \geq \sum_{i=1}^{K} \alpha_i \, \mathrm{KL}(f_X | f_{X_i}). \tag{19}$$

In order to reduce the number of free parameters, the mixture components are assumed to be the conjugates of the target distribution, i.e., $f_{X_1}$ and $f_{x_2}$ are of the same type as $f_X$, and have the means, $\bar{X}_1 = \bar{X} + \Delta\bar{X}$, and $\bar{X}_2 = \bar{X} - \Delta\bar{X}$, where $\Delta X \geq 0$. Consequently, $\alpha = 1/2$ in all the experiments, in accordance with (15).

In the case of normal distribution, there are two distribution parameters, i.e., the mean $\bar{X}$ and the variance var$[X]$. In order to account for the variance constraint in (9), the variances of the component distributions $f_{X_1}$ and $f_{X_2}$ have been equally reduced to, $p \times \mathrm{var}[X]$, $0 < p \leq 1$. The gamma distribution also has two parameters, i.e., the shape $k > 0$, and the scale $\theta > 0$. Given the scale $\theta$, the shapes of the two component distributions $f_{X_1}$ and $f_{X_2}$ are set to, $k_{1,2} = (\bar{X} \pm \Delta X)/\theta$, respectively. The normal-exponential distribution (or, exponentially-modified normal distribution) is described by three parameters, i.e., the mean and the variance of the normal distribution and the rate of the exponential distribution. The variance of the normal distribution of both components $f_{X_1}$ and $f_{X_2}$ was reduced to $0.9\mathrm{var}[X]$, and the variance of the exponential distribution was left unchanged.

The KL values for all three distributions considered are shown in a log-scale in Figure 3. It is observed that the approximations $f_{\tilde{X}}$ of $f_X$ are visually accurate for the log-KL values below $10^{-2}$. Hence, the mixture approximation $f_{\tilde{X}}$ is rather accurate for some parameter values of the target distribution, and mainly for smaller displacements $\Delta X$, as expected.

Next, we investigate the achievable magnitudes of the correlation coefficient between the random variables $X$ and $Y$ which are both generated by the mixture approximations (8). The same three marginal distributions are considered, i.e., normal, gamma, and normal-exponential distributions with the same parameters as in Figure 3. Here, the benefit of defining the bivariate distributions as the mixtures (7) to generate correlated random samples becomes evident. In particular, with the probability $\alpha$, the distributions $f_{X_1}$ and $f_{Y_1}$ are sampled independently, and with the probability $(1 - \alpha)$, the samples $X$ and $Y$ are independently generated from the distributions $f_{X_2}$ and $f_{Y_2}$, respectively. Thus, the correlated bivariate samples are generated by independently sampling from the four univariate distributions $f_{X_1}$, $f_{X_2}$, $f_{Y_1}$, and $f_{Y_2}$. The generation of normal samples is trivial, and readily available in many numerical software packages. For gamma distribution, the generator of gamma samples is either available (e.g., in Matlab as function gamrnd), or the gamma random number generator can be constructed [16]. Finally, the normal-exponential distributed samples are simply the sum of the two underlying distributions.

The achievable correlation coefficient has been measured empirically for $10^5$ bivariate random samples. The results are shown in Figure 4. The curves are in a good agreement with the theoretical values given by expressions (11). More importantly, the following conclusion can be drawn from comparing the corresponding results in Figures 3 and 4. The accurate approximation of the marginal distributions by the proposed generative mixture model limits the achievable values of the correlation coefficient to about 0.2 or 0.3, depending on the specific distributions and their parameters considered.

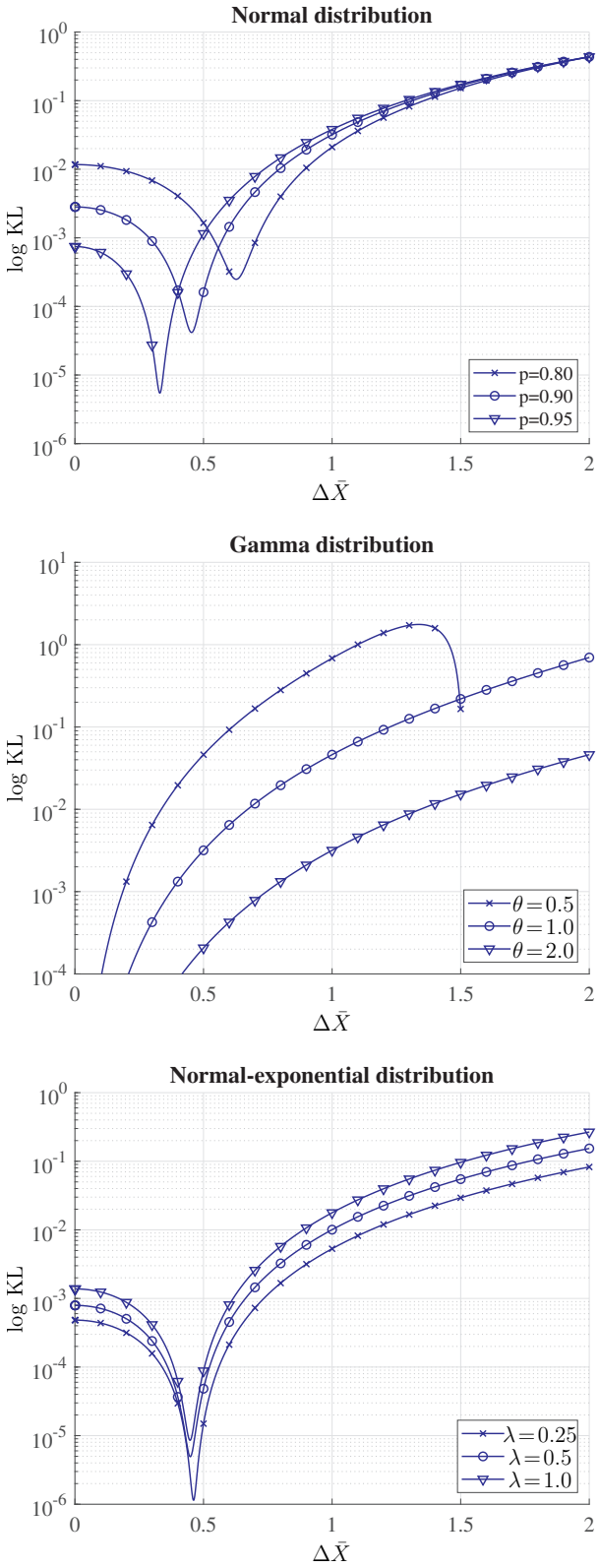


**Figure 3.** The Kullback–Leibler divergence of approximating the named univariate distributions by the two-component mixture distributions as a function of the component distributions displacement.

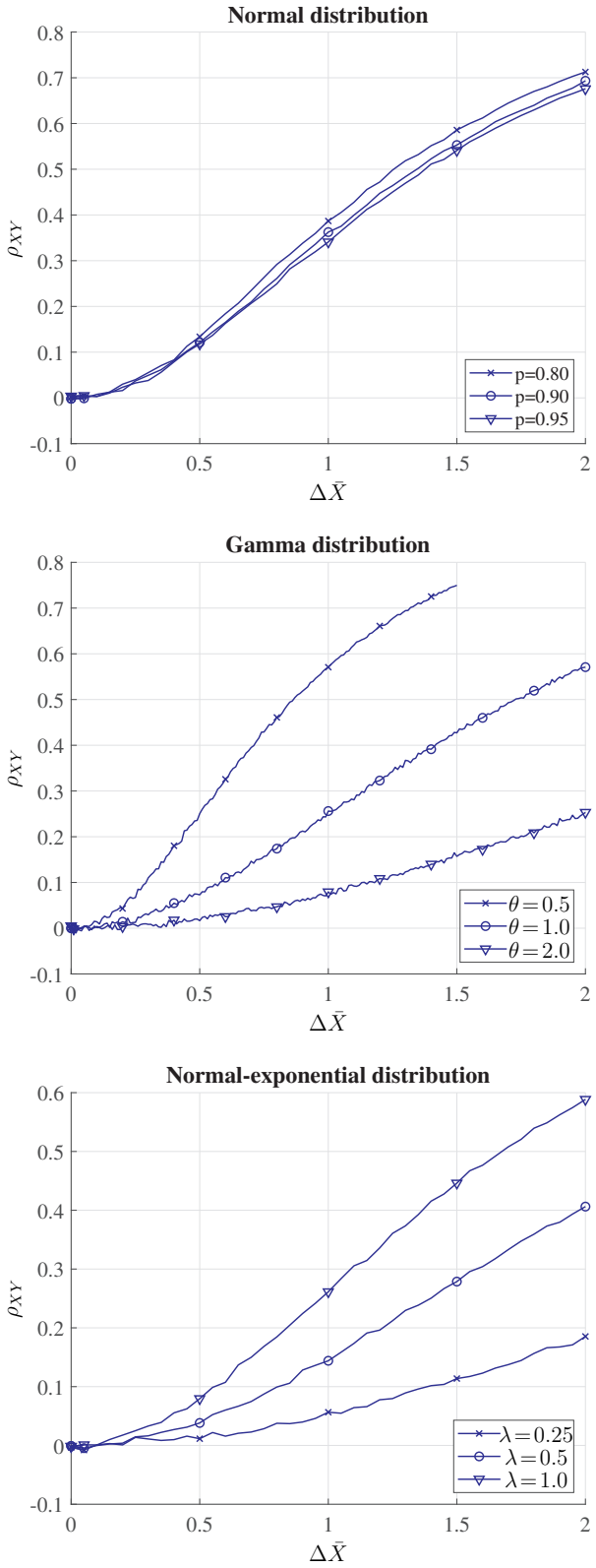


**Figure 4.** The correlation coefficient of the named bivariate distributions as a function of the component distributions displacement.

## 5. Discussion

The full statistical description of multiple time series may be difficult or impossible to obtain from a limited number of the observations. The incomplete statistics give raise to

interesting signal processing problems involving graph signals. The graph signals can be more formally defined as follows.

**Definition 1.** *The graph signal is a set of random variables representing space-time observations of stochastic processes or phenomena, for which some marginal or conditional distributions and some statistical moments are known. However, the statistical description is incomplete in the sense that a full joint distribution of all random variables cannot be uniquely determined from prior knowledge and the available observations.*

Provided that graph signals are represented as random vectors or matrices, many techniques of statistical signal processing or even machine learning can be used. For instance, Bayesian inference of parameters and unobserved model states requires a full statistical description of observed samples as the joint density or conditional density. Learning the model may require generating enough labeled data, which are consistent with the observations. Knowledge of the joint distribution is also important in controllability and observability of stochastic systems, for instance, to make optimum decisions under uncertainty. In these cases, constructing a generative model of graph observations and graph responses is crucial.

In this paper, the generative model of multiple random observations is constrained by the marginal distributions and the second order statistics. This does not uniquely define the corresponding joint distribution, but it can be further constrained by other higher order moments [7,8]. The behavior of the higher and lower order moments is described by Hölder's inequality [17]. An open research problem is whether there always exists at least one multivariate distribution given the set of marginal distributions and the pairwise correlations or covariances; it is guaranteed to exist if the observations have been obtained from a real-world system or a simulated model. There are multivariate distributions such as the multivariate Cauchy distribution for which the correlations cannot be defined. In addition, the marginal or conditional distributions may have some of their parameters undefined which increases the number of degrees of freedom. The unknown parameters could be then estimated from the known moments or other known statistics. Furthermore, practical problems often require generating the samples that are correlated in more than one dimension.

In general, there is a tradeoff between accurately approximating the marginals as mixture distributions and the achievable magnitude of the correlation coefficient as indicated by Equation (11). This has been observed in numerical examples involving two random variables assuming conjugate components in the mixture approximations of marginal distributions. The main advantage of the proposed generative model is that it can be readily sampled. The accuracy–correlation tradeoff could be improved by assuming non-conjugate mixture component distributions or by considering other types of generative models, albeit at the cost of the sampling efficacy. Alternatively, the pairwise covariance of samples $X_i$ and $Y_i$ can be reduced by simple averaging. In particular, let $\bar{X} = \sum_{i=1}^{N_1} X_i$, and, $\bar{Y} = \sum_{i=1}^{N_2} Y_i$, and assume that the samples $X_i$ and $Y_i$ are otherwise uncorrelated. Thus, let $\mathrm{E}[X_i X_j] = \mathrm{E}[Y_i Y_j] = \mathrm{E}[X_i Y_j] = 0$, for $i \neq j$, whereas $\mathrm{E}[X_i Y_i] \neq 0$. It is then straightforward to show that the correlation coefficient defined in (11) changes to

$$\rho_{\bar{X}\bar{Y}} = \sqrt{\frac{N_2}{N_1}} \rho_{XY}, \; N_2 \leq N_1. \tag{20}$$

Another strategy worth exploring is to investigate the kernel approximations of multivariate densities [11] and also the bounds of these densities. For instance, for any sets $A$ and $B$, the joint probability can bounded as

$$\Pr(A) + \Pr(B) - 1 \leq \Pr(A \cap B) \leq \sqrt{\Pr(A) \Pr(B)}. \tag{21}$$

Then, for $A = \{X : X \leq x\}$ and $B = \{Y : Y \leq y\}$, the joint cumulative function $F_{XY}$ can be bounded as

$$F_X(x) + F_Y(y) - 1 \leq F_{XY}(x, y) \leq \sqrt{F_X(x) F_Y(y)}. \tag{22}$$

For $N = 2$, the bivariate joint cumulative density is obtained by integrating Equation (2), i.e.,

$$F_{XY}(x, y) = \sum_{k=1}^{K} \alpha_k F_{X_k}(x) F_{Y_k}(y). \tag{23}$$

Assuming Equation (22), it can be bounded as

$$\sum_{k=1}^{K} \alpha_k \left(F_{X_k}(x) + F_{Y_k}(y) - 1\right) \leq \sum_{k=1}^{K} \alpha_k F_{X_k}(x) F_{Y_k}(y) \leq \sqrt{\sum_{k=1}^{K} \alpha_k F_{X_k}(x) \sum_{k=1}^{K} \alpha_k F_{Y_k}(y)}. \tag{24}$$

Moreover, in many stochastic systems, the densities evolve in time, so the discrete mixture (1) could be rewritten for the case of continuous time $t$ as

$$f(\boldsymbol{X}) = \int_{\mathcal{R}} \alpha(t)\, \tilde{f}(\boldsymbol{X}; t)\, \mathrm{d}t, \tag{25}$$

where $\alpha(t)$ is another probability distribution, i.e., $\alpha(t) \geq 0$ and $\int_{\mathcal{R}} \alpha(t)\, \mathrm{d}t = 1$. The expression (25) then represents a mean-time density of the system response.

Lastly, the generative model constructed in this paper exactly fits the first and the second statistical moments whilst approximating the marginal distributions. An alternative strategy to construct a generative model may instead emphasize fitting exactly the marginal distributions and relaxing the constraints involving the statistical moments.

## 6. Conclusions

The graph signals were defined as random vectors with incomplete knowledge of their statistics. A generative probabilistic model was then proposed to sample graph signals from given marginal distributions with given pairwise correlations. The generative model approximated the multivariate distributions by a mixture of independent univariate densities, which were then sampled to generate the correlated random sequences, which were consistent with the observations. The numerical results were presented for a bivariate case of three specific marginal distributions. The results confirmed that the proposed generative model experiences a tradeoff between accurately approximating the marginal densities and the achievable correlations, with the correlation coefficient magnitudes not greater than about 0.3. However, the cross-correlations of the observed samples can be reduced by simple averaging. Future work will focus on improving the approximation–correlation tradeoff and on defining generative models with estimation of unknown model parameters.

**Funding:** This research was funded by a start-up research grant provided by Zhejiang University.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

1. Dong, X.; Thanou, D.; Rabbat, M.; Frossard, P. Learning Graphs From Data. *IEEE Signal Process. Mag.* **2019**, *36*, 44–63. [CrossRef]
2. Pearl, J. Causal inference in statistics: An overview. *Stat. Surv.* **2009**, *3*, 96–146. [CrossRef]
3. Kay, S.M. *Fundamentals of Statistical Signal Processing: Estimation Theory*; Prentice Hall: Upper Saddle River, NJ, USA, 1993; Volume I.

4. Ortega, A.; Frossard, P.; Kovačević, J.; Moura, J.M.F.; Vandergheynst, P. Graph Signal Processing: Overview, Challenges, and Applications. *IEEE Proc.* **2018**, *106*, 808–828. [CrossRef]
5. Kotz, S.; Balakrishnan, N.; Johnson, N.L. *Continuous Multivariate Distributions*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2000.
6. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2014.
7. Foldnes, N.; Olsson, U.H. A simple simulation technique for non-normal data with pre-specified skewness, kurtosis and covariance matrix. *Multivar. Behav. Res.* **2016**, *51*, 207–219. [CrossRef] [PubMed]
8. Lyhagen, J. *Communications in Statistics-Simulation and Computation*; Taylor & Francis: Abingdon, UK, 2008; Volume 37, Chapter A Method to Generate Multivariate Data with the Desired Moments, pp. 2063–2075. [CrossRef]
9. Qu, W.; Liu, H.; Zhang, Z. A method of generating multivariate non-normal random numbers with desired multivariate skewness and kurtosis. *Behav. Res. Methods* **2020**, *52*, 939–946. [CrossRef] [PubMed]
10. Loskot, P. Polynomial Representations of High-Dimensional Observations of Random Processes. *Mathematics* **2021**, *9*, 123. [CrossRef]
11. Węglarczyk, S. Kernel density estimation and its application. *XLVIII Semin. Appl. Math.* **2018**, *23*, 00037. [CrossRef]
12. Botev, Z.I.; Grotowski, J.F.; Kroese, D.P. Kernel Density Estimation via Diffusion. *Ann. Stat.* **2010**, *38*, 2916–2957. [CrossRef]
13. O'Brien, T.A.; Kashinath, K.; Cavanaugh, N.R.; Collins, W.D.; O'Brien, J.P. A fast and objective multidimensional kernel density estimation method: FastKDE. *Comput. Stat. Data Anal.* **2016**, *101*, 148–160. [CrossRef]
14. Papoulis, A.; Pillai, S.U. *Probability, Random Variables, and Stochastic Processes*, 4th ed.; McGraw-Hill: New York, NY, USA, 2002.
15. Thede, L. *Practical Analog and Digital Filter Design*; Artech House Inc.: Norwood, MA, USA, 2004.
16. Kundu, D.; Gupta, R.D. A convenient way of generating gamma random variables using generalized exponential distribution. *Comput. Stat. Data Anal.* **2007**, *51*, 2796–2802. [CrossRef]
17. Beckenbach, E.; Bellman, R. *Inequalities*; Springer: Berlin/Heidelberg, Germany, 1961.