Article

Polynomial Representations of High-Dimensional Observations of Random Processes

Pavel Loskot (ID)

Zhejiang University/University of Illinois at Urbana-Champaign Institute, Haining 314400, China;
pavelloskot@intl.zju.edu.cn

Abstract: The paper investigates the problem of performing a correlation analysis when the number of observations is large. In such a case, it is often necessary to combine random observations to achieve dimensionality reduction of the problem. A novel class of statistical measures is obtained by approximating the Taylor expansion of a general multivariate scalar symmetric function by a univariate polynomial in the variable given as a simple sum of the original random variables. The mean value of the polynomial is then a weighted sum of statistical central sum-moments with the weights being application dependent. Computing the sum-moments is computationally efficient and amenable to mathematical analysis, provided that the distribution of the sum of random variables can be obtained. Among several auxiliary results also obtained, the first order sum-moments corresponding to sample means are used to reduce the numerical complexity of linear regression by partitioning the data into disjoint subsets. Illustrative examples provided assume the first and the second order Markov processes.

Keywords: least squares; linear regression; Markov process; moment method; multivariate function; Taylor expansion

1. Introduction

The interest in developing and improving statistical methods and models is driven by the ever increasing volumes and variety of data. Making sense of data often requires one to uncover the existing patterns and relationships in data. One of the most commonly used universal tools for data exploration is evaluation of cross-correlation among different data sets and of autocorrelation of data sequence with itself. The correlation values can be utilized to carry out sensitivity analysis, to forecast future values, to visualize complex relationships among subsystems, and to evaluate other important spatio-temporal statistical properties. However, it has been well established that correlations do not imply causalization.

The correlations measure linear dependencies between pairs of random variables which is implicitly exploited in linear regression. The correlation values rely on estimated or empirical statistical moments such as sample mean and sample variance. If the mean values are removed from data, the correlations are referred to as covariances. The pairwise correlations can be represented as a fully connected graph with edges parameterized by time shifts and possibly by amplitude adjustments between the corresponding data sequences. However, the graph may become excessively complex when the number of random variables considered is large, for example, when analyzing multiple long sequences of random observations.

The problem of extending the notion of correlations and covariances to more than two random variables has been considered previously in literature. In particular, the multirelation has been defined as a measure of linearity for multiple random variables in [1]. This measure is based on a geometric analysis of linear regression. Under the assumption of multivariate Student-t distribution, the statistical significance of the multirelation coefficient is defined in terms of eigenvalues of the correlation matrix in [2]. An univariate

correlation measure for multiple random variables is defined in [3] to be a sum of elements on the main diagonal of the covariance matrix. Linear regression is again utilized in [4] to define the multiple correlation coefficient. It is derived from the coefficient of determination of linear regression as a proportion of the variance in the dependent variable which is predictable from the independent variables.

The distance correlation measure between two random vectors based on the hypothesis testing of independence of random variables is proposed in [5]. Different operations on random variables including ordering, weighting, and a nonlinear transformation are assumed in [6] to define a family of Minkowski distance measures for random vectors. A maximal correlation measure for random vectors is introduced in [7]. It generalizes the concept of maximal information coefficient while nonlinear transformations are also used to allow assessment of nonlinear correlations. Similarly to [3], the sample cross-correlation between multiple realizations of two random vectors is shown in [8] to be proportional to the sum of diagonal elements of the product of the corresponding correlation matrices. The reference [9] investigates two-dimensional general and central moments for random matrices which are invariant to similarity transformations. Two new distance measures for random vectors are defined in [10] assuming the joint characteristic function. These measures can be also used to test for statistical independence of random vectors. The most recent paper [11] derives a linear correlation measure for multiple random variables using determinants of the correlation submatrices.

This brief literature survey indicates that there are still no commonly accepted correlation measures for multiple random variables. The measures considered in the literature are either rigorously derived but mathematically rather complicated, or there are many modifications of the existing, simpler, but well understood measures. In this paper, it is shown that by constraining the complexity of multivariate Taylor series to reduce the number of its parameters or degrees-of-freedom, the Taylor series can be rewritten as a finite degree univariate polynomial. The independent variable of this polynomial is a simple sum of the random variables considered. The polynomial coefficients are real-valued constants, and they are application dependent. The polynomial defines a many-to-one transformation of multiple random variables to another scalar random variable. The mean value of the polynomial represents a broad class of polynomial measures which can be used for any number of random variables. The mean value of each polynomial element corresponds to a general or central moment of the sum of random variables. Therefore, these moments are referred to here as sum-moments. In case of multiple random vectors, similarly to computing the correlation or covariance matrix by first concatenating the vectors into one long vector, the sum-moments can be readily defined and computed for such a concatenated random vector. The main advantages of assuming sum-moments to study statistical properties of multiple random variables or multiple random vectors are the clarity in understanding their statistical significance, and mathematical simplicity of their definition. Moreover, as long as the distribution of the sum of random variables can be found, the closed-form expression for the sum-moments can be obtained.

Before introducing the polynomial representations of random vectors in Section 4 together with central and general sum-moments, a number of auxiliary results are presented in Sections 2 and 3. In particular, Section 2 summarizes the key results and concepts from the literature concerning stationary random processes, their parameter estimation via linear regression and method of moments, and how to generate the 1st and the 2nd order Markov processes is also given. Section 3 extends the results from Section 2 by deriving additional results which are used in Sections 4 and 5 such as a low complexity approximation of linear regression and a procedure to generate multiple Gaussian processes with defined autocorrelation and cross-correlation. The main results of the paper are obtained in Section 4 including defining a class of polynomial statistical measures and sum-moments for multiple random variables and random processes. Other related concepts involving sums of random variables are also reviewed. Section 5 provides several examples to illustrate and evaluate the results obtained in the paper. The paper is concluded in Section 6.

2. Background

This section reviews key concepts and results from the literature which are used to develop new results in the subsequent sections. Specifically, the following concepts are briefly summarized: stationarity of random processes, estimation of general and central moments and of correlation and covariance using the method of moments, definition of cosine similarity and Minkowski distance, parameter estimation via linear regression, generation of the 1st and the 2nd order Markov processes, and selected properties of polynomials and multivariate functions are also given. Note that more straightforward proofs for some lemmas are only indicated and not fully and rigorously elaborated.

2.1. Random Processes

Consider a real-valued one-dimensional random process $x(t) \in \mathcal{R}$ over a continuous time $t \in \mathcal{R}$. The process is observed at N discrete time instances $t_1 < t_2 < \ldots < t_N$ corresponding to N random variables $X_i = x(t_i)$, $i = 1, \ldots, N$. The random variables $X = \{X_1, \ldots, X_N\} \in \mathcal{R}^N$ are completely statistically described by their joint density function $f_x(X)$, so that $f_x(X) \geq 0$ and $\int_{\mathcal{R}^N} f_x(X) \, dX = 1$. The process $x(t)$ is further assumed to be K-th order stationary [12].

Definition 1. *A random process $x(t)$ is K-th order stationary, if,*

$$f_x(X_1, \ldots, X_K; t_1, \ldots, t_K) = f_x(X_1, \ldots, X_K; t_1 + t_0, \ldots, t_K + t_0) \quad \forall t_0 \in \mathcal{R}.$$

Lemma 1. *The K-th order stationary process is also $(K - k)$-th order stationary, $k = 1, 2, \ldots, K - 1$, for any subset $\{X_{i_1}, \ldots, X_{i_{K-k}}\} \subseteq \{X_1, \ldots, X_N\}$.*

Proof. The unwanted random variables can be integrated out from the joint density. □

Unless otherwise stated, in the sequel, all observations of random processes are assumed to be stationary.

The expectation, $E[X] = \int_{\mathcal{R}} x f_x(x) \, dx$, of a random variable X is a measure of its mean value. A linear correlation between two random variables $x(t_1)$ and $x(t_2)$ is defined as [12],

$$R_x(t_1, t_2) = \text{corr}[x(t_1)x(t_2)] = R_x(t_2 - t_1) = R_x(\tau).$$

The (auto-) covariance measures a linear dependency between two zero-mean random variables, i.e.,

$$C_x(t_1, t_2) = \text{cov}[x(t_1), x(t_2)] = E[(x(t_1) - E[x(t_1)])(x(t_2) - E[x(t_2)])] = C_x(t_2 - t_1).$$

It can be shown that the maximum of $C_x(\tau)$, $\tau = t_2 - t_1$, occurs for $\tau = 0$, corresponding to the variance of a stationary process $x(t)$, i.e.,

$$C_x(0) = \text{var}[x(t)] = E\left[(x(t) - E[x(t)])^2\right] \quad \forall t \in \mathcal{R}.$$

The covariance $C_x(\tau)$ can be normalized, so that, $-1 \leq C_x(\tau)/C_x(0) \leq 1$. Furthermore, for real-valued regular processes, the covariance has an even symmetry, i.e., $C_x(\tau) = C_x(-\tau)$, [12].

For $N > 2$, it is convenient to define a random vector, $X = [X_1, \ldots, X_N]^T$, where $(\cdot)^T$ denotes the transpose, and the corresponding mean vector, $\bar{X} = E[X]$. Then, the covariance matrix,

$$C_x = E\left[(X - \bar{X})(X - \bar{X})^T\right] \subseteq \mathcal{R}^{N \times N}$$

has as its elements the pairwise covariances, $[C_x]_{ij} = C_x(t_j - t_i)$, $i, j = 1, 2, \ldots, N$.

Assume now the case of two *K*-th order stationary random processes $x_1(t)$ and $x_2(t)$, and the discrete time observations,

$$X_{1i} = x(t_{1i}), \quad i = 1, 2, \ldots, N_1$$
$$X_{2i} = x(t_{2i}), \quad i = 1, 2, \ldots, N_2.$$

Using the set notation, $\{X_i\}_{1:K} = \{X_1, X_2, \ldots, X_K\}$, the jointly or mutually stationary processes imply time-shift invariance of their joint density function.

Definition 2. *Random processes $x_1(t)$ and $x_2(t)$ are K-th order jointly stationary, if,*

$$f_x(\{X_{1i}\}_{1:K}, \{X_{2i}\}_{1:K}; \{t_{1i}\}_{1:K}, \{t_{2i}\}_{1:K}) = f_x(\{X_{1i}\}_{1:K}, \{X_{2i}\}_{1:K}; \{t_{1i} + t_0\}_{1:K}, \{t_{2i} + t_0\}_{1:K}) \quad \forall t_0 \in \mathcal{R}$$

is satisfied for all subsets, $\{X_{1i}\}_{1:K} \subseteq \{X_{1i}\}_{1:N_1}$, $K \leq N_1$, and $\{X_{2i}\}_{1:K} \subseteq \{X_{2i}\}_{1:N_2}$, $K \leq N_2$.

Lemma 2. *The K-th order joint stationarity implies the joint stationarity of all orders smaller than K.*

Proof. The claim follows from marginalization of the joint density function to remove unwanted variables. □

The cross-covariance of random variables $X_1 = x_1(t_1)$ and $X_2 = x_2(t_2)$ being discrete time observations of jointly stationary random processes $x_1(t)$ and $x_2(t)$, respectively, is defined as,

$$C_{x_1 x_2}(t_1, t_2) = \text{cov}[X_1, X_2] = \text{cov}[(x(t_1) - \text{E}[x(t_1)])(x(t_2) - \text{E}[x(t_2)])] = C_{x_1 x_2}(t_2 - t_1).$$

The cross-covariance can be again normalized, so it is bounded as,

$$-1 \leq C_{x_1 x_2}(\tau) / \sqrt{\text{var}[x_1(t)]\text{var}[x_2(t)]} \leq 1.$$

Note that, unlike for autocovariance, the maximum of $C_{x_1 x_2}(\tau)$ can occur for any value of the argument τ.

The covariance matrix for the random vectors $\mathbf{X}_1 = [X_{1i}, \ldots, X_{1N_1}]^T$ and $\mathbf{X}_2 = [X_{2i}, \ldots, X_{2N_2}]^T$ having the means, $\bar{\mathbf{X}}_1 = \text{E}[\mathbf{X}_1]$ and $\bar{\mathbf{X}}_2 = \text{E}[\mathbf{X}_2]$, respectively, is computed as,

$$\mathbf{C}_{x_1 x_2} = \text{E}\left[(\mathbf{X}_1 - \bar{\mathbf{X}}_1)(\mathbf{X}_2 - \bar{\mathbf{X}}_2)^T\right] \subseteq \mathcal{R}^{N_1 \times N_2}.$$

Its elements are the covariances, $[\mathbf{C}_{x_1 x_2}]_{ij} = C_{x_1 x_2}(t_{1i} - t_{2j})$, $i = 1, \ldots, N_1$, $j = 1, \ldots, N_2$.

In addition to the first order (mean value) and the second order (covariance) statistical moments, higher order statistics of a random variable X are given by the general and the central moments defined, respectively, as [13],

$$g_m(X) = \text{E}[|X|^m] \text{ and } \mu_m(X) = \text{E}[|X - \text{E}[X]|^m], \quad m = 1, 2, \ldots \tag{1}$$

where $|X|$ denotes the absolute value of scalar variable X. The positive integer-valued moments (1) facilitate mathematically tractable integration, and prevent producing complex numbers, if the argument is negative. Note also that the absolute value in (1) is necessary if X is complex valued, or if m is odd, in order to make the moments to be real-valued and convex. The central moment can be normalized by the variance as,

$$\mu_m(X) = \text{E}[|X - \text{E}[X]|^m] / \text{E}\left[(X - \text{E}[X])^2\right]^{m/2}.$$

The cosine similarity between two equal-length vectors, $\boldsymbol{X}_1 = [X_{11}, \ldots, X_{1N}]^T$ and $\boldsymbol{X}_2 = [X_{21}, \ldots, X_{2N}]^T$, is defined as [14],

$$S_{\cos}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{\sum_{i=1}^{N} X_{1i} X_{2i}}{\sqrt{\sum_{i=1}^{N} X_{1i}^2} \sqrt{\sum_{i=1}^{N} X_{2i}^2}}.$$

The Minkowski distance between two equal-length vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is defined as the l_m-norm [6], i.e.,

$$S_{\text{mnk}}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_m = \Big(\sum_{i=1}^{N} \Big| X_{1i} - X_{2i} \Big|^m \Big)^{1/m}, \quad m = 1, 2, \ldots$$

2.2. Estimation Methods

Statistical moments can be empirically estimated from measured data using sample moments. Such inference strategy is referred to as the method of moments [15]. In particular, under the ergodicity assumption, a natural estimator of the mean value of a random variable X from its measurements X_i is the sample moment,

$$\bar{X} = \mathrm{E}[X] \approx \frac{1}{N} \sum_{i=1}^{N} X_i.$$

The sample mean estimator is unbiased and consistent [15]. More generally, the sample mean estimator of the first moment of a transformed random variable $h(X)$ is,

$$\mathrm{E}[h(X)] \approx \frac{1}{N-d} \sum_{i=1}^{N} h(X_i)$$

where d is the number of degrees of freedom used in the transformation h, i.e., the number of other parameters which must be estimated. For example, the variance of X is estimated as,

$$\mathrm{var}[X] = \mathrm{E}\Big[|X - \bar{X}|^2 \Big] \approx \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \hat{\bar{X}})^2$$

where $\hat{\bar{X}}$ is the estimate of the mean value $\bar{X}$ of X.

Assuming random sequences $\{X_{1i}\}_{1:N_1}$ and $\{X_{2i}\}_{1:N_2}$, their autocovariance and cross-covariance, respectively, are estimated as [15],

$$C_x(k) = \mathrm{E}[X_i X_{i+k}] \approx \frac{1}{N-k} \sum_{i=1}^{N-k} X_i X_{i+k}, \quad k \ll N.$$

$$C_{x_1 x_2}(k) = \mathrm{E}\Big[X_{1i} X_{2(i+k)} \Big] \approx \frac{1}{N-k} \sum_{i=1}^{N-k} X_{1i} X_{2(i+k)}, \quad k \ll N = \min(N_1, N_2).$$

Since these estimators are consistent, the condition $k \ll N$ is necessary to combine a sufficient number of samples and achieve an acceptable estimation accuracy.

The parameters of a random process can be estimated by fitting a suitable data model to the measurements X_i. Denote such data model as, $\mathrm{E}[X_i] = M_i(\boldsymbol{P})$, $i = 1, 2, \ldots, N$. Assuming the least-squares (LS) criterion, the vector of parameters, $\boldsymbol{P} = [P_1, \ldots, P_D]^T \subseteq \mathcal{R}^D$, is estimated as,

$$\hat{\boldsymbol{P}} = \underset{\boldsymbol{P}}{\arg\min} \sum_{i=1}^{N} (X_i - M_i(\boldsymbol{P}))^2.$$

For continuous parameters, the minimum is obtained using the derivatives, i.e., let $\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{P}} M_i(\boldsymbol{P}) = \dot{M}_i(\boldsymbol{P})$, and,

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{P}} \sum_{i=1}^{N} (X_i - M_i(\boldsymbol{P}))^2 \overset{!}{=} 0 \quad \Leftrightarrow \quad \sum_{i=1}^{N} \dot{M}_i(\hat{\boldsymbol{P}}) M_i(\hat{\boldsymbol{P}}) = \sum_{i=1}^{N} \dot{M}_i(\hat{\boldsymbol{P}}) X_i.$$

For a linear data model, $M_i(\boldsymbol{P}) = \boldsymbol{w}_i^T \boldsymbol{P}$ where $\boldsymbol{w}_i = [w_{1i}, \ldots, w_{Di}]^T$ are known coefficients, the LS estimate can be obtained in the closed-form, i.e.,

$$\hat{\boldsymbol{P}} = \left(\sum_{i=1}^{N} \boldsymbol{w}_i \boldsymbol{w}_i^T \right)^{-1} \sum_{i=1}^{N} \boldsymbol{w}_i X_i \tag{2}$$

where $(\cdot)^{-1}$ denotes the matrix inverse.

2.3. Generating Random Processes

The task is to generate a discrete time stationary random process with a given probability density and a given autocovariance. The usual strategy is to generate a correlated Gaussian process followed by a nonlinear memoryless transformation. For instance, the autoregressive (AR) process described by the second order difference equation with constant coefficients a_1, a_2, and $b > 0$ [12], i.e.,

$$x(n) + a_1 x(n-1) + a_2 x(n-2) = b\, u(n)$$

generates the 2nd order Markov process from a zero-mean white (i.e., uncorrelated) process $u(n)$. For $a_1 = -2\sqrt{\frac{1-a}{1+a}}$ and $a_2 = \frac{1-a}{1+a}$, $0 < a < 1$, this process has the autocovariance,

$$C_{2\mathrm{MP}}(k) = b^2 \underbrace{\frac{(1+a)^2}{4a^3}}_{\sigma^2} \underbrace{\left(\frac{1-a}{1+a}\right)^{|k|/2}}_{\doteq (1-a)^{|k|}} (1 + a|k|)$$

$$\doteq \sigma^2 (1-a)^{|k|} (1 + a\,|k|) = \sigma^2 \mathrm{e}^{-|k|(-\ln(1-a))} (1 + a\,|k|). \tag{3}$$

On the other hand, for $a_1 = a > 0$, and $a_2 = 0$, the AR process,

$$x(n) + ax(n-1) = b\, u(n)$$

generates the 1st order Markov process with the autocovariance,

$$C_{1\mathrm{MP}}(k) = \underbrace{\frac{b^2}{1-a^2}}_{\sigma^2} a^{|k|} = \sigma^2 \mathrm{e}^{-\alpha|k|}, \quad a = \mathrm{e}^{-\alpha}. \tag{4}$$

Lemma 3. *[16] The stationary random process $x(k)$ with autocovariance $C_x(k)$ is transformed by a linear time-invariant system with real-valued impulse response $h(k)$ into another stationary random process $y(k) = \sum_{i=-\infty}^{\infty} h(i)x(k-i) \equiv h(k) \circledast x(k)$ with autocovariance, $C_y(k) = h(k) \circledast h(-k) \circledast C_x(k)$. The symbol, $\circledast$, denotes (discrete time) convolution.*

Proof. By definition, the output covariance, $C_y(k, l) = \mathrm{E}[(y(k) - \mathrm{E}[y(k)])(y(l) - \mathrm{E}[y(l)])]$. Substituting $y(k) = \sum_{i=-\infty}^{\infty} h(i)x(k-i)$, and rearranging, the covariance, $C_y(k-l) = \sum_i \sum_m h(m)C_x(i-m)h(k-i) = h(k) \circledast h(-k) \circledast C_x(k)$. $\square$

Lemma 4. *A stationary random process at the output of a linear or nonlinear time-invariant system remains stationary.*

Proof. For any multivariate function $h(x) \in \mathcal{R}$ and any $t_0 \in \mathcal{R}$, the expectation,

$$\mathrm{E}[h(x)] = \int_{\mathcal{R}^N} h(x) f_X(x; t_1, \ldots, t_N) \, \mathrm{d}x = \int_{\mathcal{R}^N} h(x) f_X(x; t_1 - t_0, \ldots, t_N - t_0) \, \mathrm{d}x$$

assuming Definition 1 and provided that dimension N of observations is at most equal to the order of stationarity K. $\square$

For shorter sequences, the linear transformation, $X = TU$, can be used to generate a normally distributed vector $X \in \mathcal{R}^N$ having the covariance matrix, $C_x = TT^T$, from uncorrelated Gaussian vector $U \in \mathcal{R}^N$. The mean, $\mathrm{E}[X] = T\mathrm{E}[U]$. For longer sequences, a linear time-invariant filter can be equivalently used as indicated by Lemma 3.

2.4. Polynomials and Multivariate Functions

Lemma 5. *A univariate twice differentiable function $p(x)$ is convex, if and only if, $\frac{\mathrm{d}^2}{\mathrm{d}x^2} p(x) = \ddot{p}(x) > 0$ for $\forall x \in \mathcal{R}$. More generally, a twice differentiable multivariate function $f : \mathcal{R}^N \mapsto \mathcal{R}$ is convex, if and only if, the domain of f is convex, and its Hessian is positive semidefinite, i.e., $\nabla^2 f \geq 0$ for $\forall x \in \mathrm{dom}\ f$.*

Proof. See [17] (Sec. 3.1). $\square$

Consequently, convex polynomials can be generated as follows.

Lemma 6. *Let $Q \in \mathcal{R}^{m \times m}$ be a positive semidefinite matrix, and assume a polynomial, $\ddot{p}(x) = \sum_{i=0}^{2m-2} p_i x^i$ for $\forall x \in \mathcal{R}$ where $p_i = \sum_{k+l=i} Q_{kl}$. Then, for any $q_0, q_1 \in \mathcal{R}$, the polynomial $p(x)$ of degree $2m$,*

$$p(x) = q_0 + q_1 x + \sum_{i=0}^{2m-2} \frac{p_i}{(i+1)(i+2)} x^{2+i}$$

is convex.

Proof. Let $x = [x^0, x^1, \ldots, x^{m-1}]^T$. Then, $x^T Q x = \sum_{i=0}^{2m-2} p_i x^i = \ddot{p}(x) \geq 0$ for $\forall x$, since Q is positive semidefinite. Using Lemma 5 concludes the proof. $\square$

For a non-negative integer $m \in \{0\} \cup \mathbb{N}_+ = \{0, 1, 2, \ldots\}$, assume the following notations to simplify the subsequent mathematical expressions [18]:

$$
\begin{aligned}
n &= \{n_1, n_2, \ldots, n_N\} \in \{\mathbb{N}_+ \cup 0\}^N \\
x &= \{x_1, x_2, \ldots, x_N\} \in \mathcal{R}^N \\
\|x\|_p &= \left(x_1^p + x_2^p + \cdots + x_N^p \right)^{1/p} \\
\|x\|_1 &= |x_1| + |x_2| + \cdots + |x_N| \\
|x|_1 &= x_1 + x_2 + \cdots + x_N \\
h &= \{h_1, h_2, \ldots, h_N\} \in \mathcal{R}^N \\
h^n &= h_1^{n_1} h_2^{n_2} \cdots h_N^{n_N} \\
\partial^n f(x) &= \partial_1^{n_1} \partial_2^{n_2} \cdots \partial_N^{n_N} f(x) = \frac{\partial^{|n|_1}}{\partial x_1^{n_1} x_2^{n_2} \cdots x_N^{n_N}} f(x) \\
m! &= \prod_{i=1}^m i \\
n! &= n_1! \, n_2! \cdots n_N!
\end{aligned}
$$

Note that $|x|_1$ denotes the sum of elements of x whereas $\|x\|_1$ is the sum of absolute values of its elements.

Lemma 7. *The m-th power of a finite sum can be expanded as [19],*

$$|x|_1^m = \sum_{|n|_1 = m} \frac{m!}{n!} x^n.$$

Proof. See [18]. □

Theorem 1. *The multivariate Taylor's expansion of a* $(m+1)$*-order differentiable function* $f:$ $\mathcal{R}^N \mapsto \mathcal{R}$ *about the point* $\boldsymbol{x} \in \mathcal{R}^N$ *is written as [19],*

$$f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + \sum_{|\boldsymbol{n}|_1 \leq m} \frac{\partial^{\boldsymbol{n}} f(\boldsymbol{x})}{\boldsymbol{n}!} \boldsymbol{h}^{\boldsymbol{n}} + \sum_{|\boldsymbol{n}|_1 = m+1} \frac{\partial^{\boldsymbol{n}} f(\boldsymbol{x} + t\,\boldsymbol{h})}{\boldsymbol{n}!} \boldsymbol{h}^{\boldsymbol{n}}$$

for some $t \in (0,1)$.

Proof. See [18,19]. □

Definition 3. *A multivariate function* $f(\boldsymbol{x}) = f(x_1, \ldots, x_N)$ *is said to be symmetric, if, for any permutation of its arguments* $\boldsymbol{x}$ *denoted as* $\boldsymbol{x}'$, $f(\boldsymbol{x}) = f(\boldsymbol{x}')$.

3. Background Extensions

In this section, additional results are obtained which are used in the next section to introduce statistical sum-moments for random vectors. In particular, the mean cosine similarity, the mean Minkowski distance as well as the higher central moments for random vectors are defined. A polynomial approximation of univariate functions is shown to be a linear regression problem. A numerically efficient solution of the LS problem is derived. Finally, a procedure to generate multiple Gaussian processes with defined autocovariance and cross-covariance is devised.

Recall the second moment of a random variable X, i.e.,

$$\mu_2 = \mathrm{E}\Big[(X - c)^2\Big], \quad c \in \mathcal{R}.$$

It is straightforward to show that μ_2 is minimized for $c = \mathrm{E}[X]$, giving the variance of X. On the other hand, $\mu_2 = 0$, if and only if $c = \mathrm{E}[X] \pm \sqrt{-\mathrm{var}[X]}$.

For random vectors, both cosine similarity and Minkowski distance are random variables. Assuming that the vectors are jointly stationary, and their elements are identically distributed, the mean cosine similarity can be defined as,

$$\bar{S}_{\cos}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{\sum_{i=1}^{N} \mathrm{E}[X_{1i}X_{2i}]}{\sqrt{\sum_{i=1}^{N} \mathrm{E}[X_{1i}^2] - \bar{X}_{1i}^2}\sqrt{\sum_{i=1}^{N} \mathrm{E}[X_{2i}^2] - \bar{X}_{2i}^2}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathrm{E}[X_{1i}X_{2i}]}{\sigma_1 \sigma_2} = \frac{1}{N}\sum_{i=1}^{N} \rho_{1i,2i}$$

where $\sigma_1^2 = \mathrm{var}[X_{1i}]$, $\sigma_2^2 = \mathrm{var}[X_{2i}]$, and, $\rho_{1i,2i}$ denotes the Pearson correlation coefficient of the i-th elements of the vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. It should be noted that this definition of the mean cosine similarity does not account for other correlations, $\mathrm{E}[X_{1i}X_{2j}], i \neq j$.

The mean Minkowski distance for random vectors can be defined as,

$$\bar{S}_{\mathrm{mnk}}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \Big(\sum_{i=1}^{N} \mathrm{E}\big[|X_{1i} - X_{2i}|^m\big]\Big)^{1/m}, \quad m = 1, 2, \ldots \tag{5}$$

Recognizing the m-th general moment in (5), the m-th power of the mean Minkowski distance can be normalized as,

$$\tilde{S}_{\mathrm{mnk}}^m(\boldsymbol{X}_1, \boldsymbol{X}_2) = \sum_{i=1}^{N} \frac{\mathrm{E}\big[|X_{1i} - X_{2i}|^m\big]}{N\Big(\mathrm{E}\big[|X_{1i} - X_{2i}|^2\big]\Big)^{m/2}} = \bar{\mu}_m(\boldsymbol{X}_1 - \boldsymbol{X}_2), \quad m = 1, 2, \ldots$$

where the average Minkowski distance between two random vectors is,

$$\bar{\mu}_m(\boldsymbol{X}_1 - \boldsymbol{X}_2) = \frac{1}{N}\sum_{i=1}^{N} \mu_m(X_{1i} - X_{2i}), \quad m = 1, 2, \ldots$$

Furthermore, note that for $m = 2$,

$$\frac{1}{2}\left(E\left[\|\mathbf{X}_1 - \mathbf{X}_2\|_2^2\right] + E\left[\|\mathbf{X}_1 + \mathbf{X}_2\|_2^2\right]\right) = E\left[\|\mathbf{X}_1\|_2^2\right] + E\left[\|\mathbf{X}_2\|_2^2\right].$$

Assuming positive integers $\mathbf{m} = \{m_1, m_2, \ldots, m_N\}$, the higher order joint central moments of random vector $\mathbf{X} = \{X_i\}_{1:N}$ can be defined as,

$$\mu_{m_1,\ldots,m_N}(X_1,\ldots,X_N) = E\left[\prod_{i=1}^{N}(X_i - \bar{X}_i)^{m_i}\right] \tag{6}$$

or, using more compact index notation as,

$$\mu_{\mathbf{m}}(\mathbf{X}) = E[(\mathbf{X} - \bar{\mathbf{X}})^{\mathbf{m}}].$$

3.1. Linear LS Estimation

The linear LS estimation can be used to fit a degree $(D - 1)$ polynomial to N samples of a random process $x(t)$ at discrete time instances $t_1, t_2, \ldots, t_N$. Hence, consider the polynomial data model,

$$E[x(t)] \approx M(t; \mathbf{P}) = \sum_{k=1}^{D} P_k t^{k-1}.$$

Denoting $\mathbf{w}_i = [t_i^0, t_i^1, \ldots, t_i^{D-1}]^T$, the linear LS solution (2) gives the estimates,

$$\hat{\mathbf{P}} = \left(\sum_{i=1}^{N}\begin{bmatrix} t_i^0 & t_i^1 & \cdots & t_i^{D-1} \\ t_i^1 & t_i^2 & \cdots & t_i^D \\ \vdots & \vdots & \ddots & \vdots \\ t_i^{D-1} & t_i^D & \cdots & t_i^{2D-2} \end{bmatrix}\right)^{-1}\sum_{i=1}^{N}x(t_i)\begin{bmatrix} t_i^0 \\ t_i^1 \\ \vdots \\ t_i^{D-1} \end{bmatrix}.$$

Assuming $D = 2$ parameters, i.e., the linear LS regression for a straight line, the parameters P_1 and P_2 to be estimated must satisfy the following equality:

$$\frac{\mathrm{d}}{\mathrm{d}P_1}\sum_{i=1}^{N}(X_i - w_{1i}P_1 - w_{2i}P_2)^2 = -2\sum_{i=1}^{N}w_{1i}X_i + 2\sum_{i=1}^{N}w_{1i}^2 P_1 + 2\sum_{i=1}^{N}w_{1i}w_{2i}P_2 \overset{!}{=} 0.$$

Denoting the weighted averages, $\bar{X} = \sum_{i=1}^{N}w_{1i}X_i$, $\|\mathbf{w}_1\|_2^2 = \sum_{i=1}^{N}w_{1i}^2$, and $\bar{w}_{12} = \sum_{i=1}^{N}w_{1i}w_{2i}$, a necessary but not sufficient condition for the linear LS estimation of parameters P_1 and P_2 is,

$$\bar{X} = \|\mathbf{w}_1\|_2^2 P_1 + \bar{w}_{12}P_2. \tag{7}$$

In the LS terminology, the values $\|\mathbf{w}_1\|_2^2$ and $\bar{w}_{12}$ represent independent variables whereas $\bar{X}$ is a dependent variable.

Note that all N measurements are used in (7). However, if N is sufficiently large, the data points could be split into two disjoint sets of N_1 and N_2 elements, respectively, $N_1 + N_2 = N$. For convenience, denote the sums,

$$\bar{X}_1 = \frac{1}{a_1}\sum_{i \in \mathcal{I}_1}w_{1i}X_i, \quad \bar{W}_{11} = \frac{1}{a_1}\sum_{i \in \mathcal{I}_1}w_{1i}^2, \quad \bar{W}_{12} = \frac{1}{a_1}\sum_{i \in \mathcal{I}_1}w_{1i}w_{2i},$$

$$\bar{X}_2 = \frac{1}{a_2}\sum_{i \in \mathcal{I}_2}w_{1i}X_i, \quad \bar{W}_{21} = \frac{1}{a_2}\sum_{i \in \mathcal{I}_2}w_{1i}^2, \quad \bar{W}_{22} = \frac{1}{a_2}\sum_{i \in \mathcal{I}_2}w_{1i}w_{2i},$$

where $a_1, a_2 > 1$ are some constants (to be determined later), and $\mathcal{I}_1$ and $\mathcal{I}_2$ are two disjoint index sets, such that $\mathcal{I}_1 \cup \mathcal{I}_2 = \{1, 2, \ldots, N\}$, and the cardinality, $|\mathcal{I}_1| = N_1$ and $|\mathcal{I}_2| = N_2$. Note also that,

$$
\begin{aligned}
\|w_1\|_2^2 &= a_1 \bar{W}_{11} + a_2 \bar{W}_{21} \\
\bar{w}_{12} &= a_1 \bar{W}_{12} + a_2 \bar{W}_{22} \\
\bar{X} &= a_1 \bar{X}_1 + a_2 \bar{X}_2.
\end{aligned}
\tag{8}
$$

Consequently, the approximate LS estimates of parameters P_1 and P_2 are readily computed as,

$$
\begin{bmatrix} \hat{P}_1 \\ \hat{P}_2 \end{bmatrix} = \begin{bmatrix} \bar{W}_{11} & \bar{W}_{12} \\ \bar{W}_{21} & \bar{W}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \end{bmatrix}.
\tag{9}
$$

There are 2^N possibilities how to split N data points into two disjoint subsets indexed by $\mathcal{I}_1$ and $\mathcal{I}_2$. More importantly, the estimates (9) do not guarantee the minimum LS fit, i.e., achieving the minimum squared error (MSE). However, the complexity of performing the LS fit is greatly reduced by splitting the data, since independently of the value of $N \gg 1$, only a 2×2 matrix needs to be inverted. The optimum LS fit and the approximate LS fit (9) are depicted in Figure 1. The points A_1 and A_2 in Figure 1 correspond to the data subsets indexed by $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively. The mid-point, $B = a_1 A_1 + a_2 A_2$, follows from (8). Note that B is always located at the intersection of the optimum and the approximate lines LS_{opt} and LS_{apr}, respectively. The vertical arrows at points A_1 and A_2 in Figure 1 indicate that the dependent values $\bar{X}_1$ and $\bar{X}_2$ are random variables.
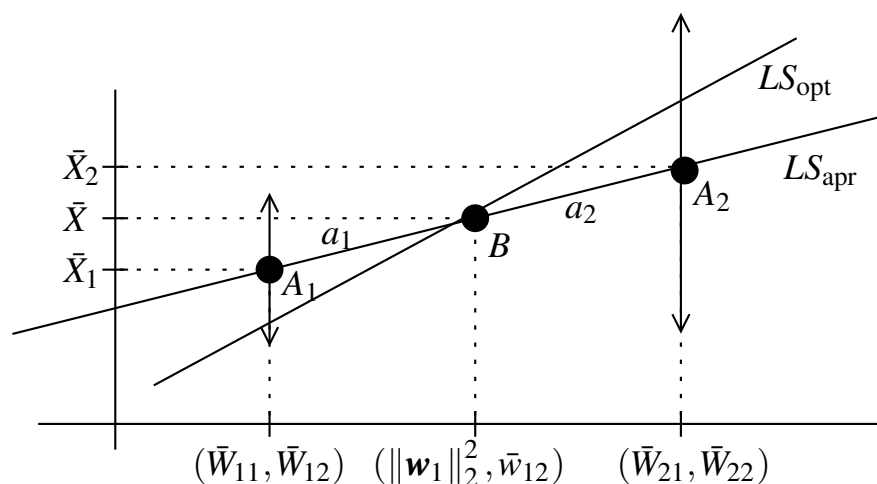

Figure 1. The exact (LS_{opt}) and the reduced complexity (LS_{apr}) linear LS regression.

The larger the variation of the gradient of the line LS_{apr} in Figure 1, the larger the uncertainty and the probability that the line LS_{apr} deviates from the optimum regression line LS_{opt}. Since the line LS_{apr} is defined by points $\bar{X}_1$ and $\bar{X}_2$, and always, $B \in LS_{\mathrm{apr}}$, the spread of random variables $\bar{X}_1$ and $\bar{X}_2$ about their mean values affect the likelihood that LS_{apr} deviates from LS_{opt}. In particular, given $0 < p < 1$, there exists $\xi > 0$, such that the probability of the gradient G_{apr} of LS_{apr} to be within the given bounds is at least,

$$
\Pr\left(b_{\mathrm{l}}(\xi) < G_{\mathrm{apr}} < b_{\mathrm{u}}(\xi)\right) \geq p
\tag{10}
$$

where

$$
\begin{aligned}
b_{\mathrm{l}}(\xi) &= \frac{\left(\mathrm{E}[\bar{X}_2] - \xi\sqrt{\mathrm{var}[\bar{X}_2]}\right) - \left(\mathrm{E}[\bar{X}_1] + \xi\sqrt{\mathrm{var}[\bar{X}_1]}\right)}{\bar{W}_{22} - \bar{W}_{12}} \\
b_{\mathrm{u}}(\xi) &= \frac{\left(\mathrm{E}[\bar{X}_2] + \xi\sqrt{\mathrm{var}[\bar{X}_2]}\right) - \left(\mathrm{E}[\bar{X}_1] - \xi\sqrt{\mathrm{var}[\bar{X}_1]}\right)}{\bar{W}_{22} - \bar{W}_{12}}.
\end{aligned}
$$

For stationary measurements (cf. (9)), the means and the variances in (10) are equal to,

$$
\begin{aligned}
\mathrm{E}[\bar{X}_1] &= \mathrm{E}[\bar{X}]N_1/(a_1N) & \mathrm{var}[\bar{X}_1] &= \mathrm{var}[\bar{X}]N_1/(a_1^2 N) \\
\mathrm{E}[\bar{X}_2] &= \mathrm{E}[\bar{X}]N_2/(a_2N) & \mathrm{var}[\bar{X}_2] &= \mathrm{var}[\bar{X}]N_2/(a_2^2 N).
\end{aligned}
$$

The uncertainty in computing G_{apr} from data, and thus, also the probability of line LS_{apr} deviating from line LS_{opt} is inversely proportional to the width of the interval in (10), i.e.,

$$
b_{\mathrm{u}}(\xi) - b_{\mathrm{l}}(\xi) = \frac{2\xi\left(\sqrt{\mathrm{var}[\bar{X}_1]} + \sqrt{\mathrm{var}[\bar{X}_2]}\right)}{\bar{W}_{22} - \bar{W}_{12}} \propto \frac{\frac{\sqrt{N_1}}{a_1} + \frac{\sqrt{N_2}}{a_2}}{\bar{W}_{22} - \bar{W}_{12}}. \tag{11}
$$

Consequently, the numerator of (11) must be minimized, and the denominator maximized.

In order to minimize the numerator in (11), it is convenient to choose, $a_1 = N_1^\nu$ and $a_2 = N_2^\nu$, where $\nu \in \mathcal{R}^+$ is a constant to be optimized. It is straightforward to show that the expression, $(N_1^{1/2-\nu} + N_2^{1/2-\nu})$ is convex, i.e., it has a unique global minimum for $\forall \nu > 1/2$. Hence, a necessary condition to reduce the approximation error is that $a_1 > \sqrt{N_1}$ and $a_2 > \sqrt{N_2}$. For numerical convenience, let $a_1 = N_1$ and $a_2 = N_2$. Then, the dependent and independent variables assumed in (9) become arithmetic averages, and the optimum index subsets have the cardinality,

$$
\begin{cases}
|\mathcal{I}_1| = |\mathcal{I}_2| = N/2 & N - \text{even} \\
|\mathcal{I}_1| = |\mathcal{I}_2| \pm 1 = (N \pm 1)/2 & N - \text{odd}.
\end{cases} \tag{12}
$$

In order to maximize the denominator in (11), assume that the independent variables $(\|\boldsymbol{w}_1\|_2^2, \bar{w}_{12})$ in (9) are sorted by w_{li}, i.e., let $w_{11} < w_{12} < \ldots < w_{1N}$. This ordering and the condition (12) suggests that the disjoint index sets $\mathcal{I}_1$ and $\mathcal{I}_2$ maximizing the difference, $\bar{W}_{22} - \bar{W}_{12}$, are,

$$
\begin{cases}
\mathcal{I}_1 = \{1, 2, \ldots, N/2\}, \mathcal{I}_2 = \{N/2 + 1, \ldots, N\} & N - \text{even} \\
\mathcal{I}_1 = \{1, 2, \ldots, (N-1)/2\}, \mathcal{I}_2 = \{(N+1)/2, \ldots, N\} & N - \text{odd} \\
\text{or, } \mathcal{I}_1 = \{1, 2, \ldots, (N+1)/2\}, \mathcal{I}_2 = \{(N+3)/2, \ldots, N\}.
\end{cases}
$$

Such a partitioning corresponds to splitting the data into two equal (N-even) or approximately equal (N-odd) sized subsets by the median (2-quantile) index point.

In summary, the approximate linear LS regression can be efficiently computed with a good accuracy by splitting the data into multiple disjoint subsets, calculating the average data points in each of these subsets, and then solving the set of linear equations with the same (or smaller) number of unknown parameters. The data splitting should exploit ordering of data points by one of the independent variables. It can be expected that the accuracy of approximate LS regression is going to improve with the number of data points N. Numerical evaluation of the approximate LS regression is considered in Section 5.

3.2. Generating Pairwise-Correlated Gaussian Processes

How to generate a single correlated Gaussian process is well established in the literature, and it has been described in Section 2.3. Moreover, the linear transformation to generate correlated Gaussian variables from uncorrelated Gaussian variables does not have to be square. A sufficient condition on the rank of linear transformation to obtain a positive definite covariance matrix is given by the following lemma.

Lemma 8. *The matrix, $\boldsymbol{T}^T \boldsymbol{T}$, is positive definite, provided that the matrix, $\boldsymbol{T} \in \mathcal{R}^{N_1 \times N_2}$, has rank N_2.*

Proof. The rank N_2 of $\boldsymbol{T}$ implies that $\boldsymbol{T}$ consists of N_2 linearly independent columns and that $N_1 \geq N_2$. The matrix $\boldsymbol{T}^T \boldsymbol{T}$ is positive definite, provided that $\boldsymbol{U}^T \boldsymbol{T}^T \boldsymbol{T} \boldsymbol{U} = \|\boldsymbol{T}\boldsymbol{U}\|_2^2 >$

0 for $\forall U \in \mathcal{R}^{N_2}$ where $\|\cdot\|_2$ denotes the Euclidean norm of a vector. Since the columns of T are linearly independent, $\|TU\|_2 = 0$, if and only if $U = 0$. $\quad\square$

Corollary 1. *The matrix, TT^T, is positive definite, provided that the matrix, $T \in \mathcal{R}^{N_1 \times N_2}$, has rank N_1.*

Corollary 2. *A rank N_1 linear transformation T of uncorrelated Gaussian vector $U \in \mathcal{R}^{N_2}$ generates $N_1 \leq N_2$ correlated Gaussian variables having the (positive definite) covariance matrix, TT^T.*

Furthermore, it is often necessary to generate multiple mutually correlated Gaussian processes with given autocorrelation as well as cross-correlation.

Lemma 9. *The linear transformation,*

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} T_1 & K \\ 0 & T_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{13}$$

generates a pair of correlated Gaussian vectors $x_1 \in \mathcal{R}^{N_1}$ and $x_2 \in \mathcal{R}^{N_2}$ from uncorrelated zero-mean Gaussian vectors $u_1, u_2 \in \mathcal{R}^{N}$ where 0 denotes a zero matrix, and according to Corollary 2, it is necessary that, $\max(N_1, N_2) \leq N$. The corresponding (auto-) correlation and cross-correlation matrices are,

$$\mathrm{E}\left[u_1 u_1^T\right] = \sigma_1^2 I, \quad \mathrm{E}\left[u_2 u_2^T\right] = \sigma_2^2 I, \quad \mathrm{E}\left[u_1 u_2^T\right] = 0$$

$$C_{x_1} = \mathrm{E}\left[x_1 x_1^T\right] = T_1 T_1^T + KK^T, \ C_{x_2} = \mathrm{E}\left[x_2 x_2^T\right] = T_2 T_2^T, \ C_{x_1 x_2} = \mathrm{E}\left[x_1 x_2^T\right] = KT_2^T = C_{x_2 x_1}^T$$

where I denotes an identity matrix.

Proof. The proof is straightforward by substituting (13) to definitions of C_{x_1}, C_{x_2} and $C_{x_1 x_2}$. $\quad\square$

The following corollary details the procedure described in Lemma 9.

Corollary 3. *Given T_2, calculate the (auto-) correlation matrix, $C_{x_2} = KT_2^T$, or vice versa. Then, given the cross-correlation matrix, $C_{x_1 x_2}$, compute $K = C_{x_1 x_2} T_2 C_{x_2}^{-T}$. Finally, given T_1, calculate the (auto-) correlation matrix, $C_{x_1} = T_1 T_1^T + KK^T$, or, obtain T_1 by solving the matrix equation, $T_1 T_1^T = C_{x_1} - KK^T$. Note that the matrix equation, $T^T T = C$, can be solved for T using the singular value decomposition, $C = U \Lambda U^T$ where U is a unitary matrix and Λ is a diagonal matrix of eigenvalues. Then, $T = U\sqrt{\Lambda}$.*

4. Polynomial Statistics and Sum-Moments for Vectors of Random Variables

The main objective of this section is to define a universal function to effectively measure the statistics of random vectors and random processes observed at multiple discrete time instances. The measure function should: (1) be universally applicable for an arbitrary number of random variables and random vectors, (2) be symmetric, so that all random variables are considered equally, (3) lead to mathematically tractable expressions, and (4) be convex to allow defining convex optimization problems.

Let $f : \mathcal{R}^N \mapsto \mathcal{R}$ denote such a mapping or transformation of N random variables X_i to a scalar random variable Y, i.e.,

$$Y = f(X_1, X_2, \ldots, X_N) = f(x). \tag{14}$$

In order to satisfy the symmetry requirement, the random variables X_i can be first combined as,

$$Y = f(X_1 \circ X_2 \circ \cdots \circ X_N)$$

using a binary commutative operator, $\circ$, such as addition or multiplication. In case of addition, it is important that the function f is nonlinear. The nonlinearity can be also used to limit the extreme values of the combined variables.

For a random process $x(t)$, the random variables are defined as, $X_i = x(t_i)$. Define a vector of discrete time instances, $\boldsymbol{t} = (t_1, \ldots, t_N)$, and assume the index notation, $x(\boldsymbol{t}) = \{X_1 = x(t_1), \ldots, X_N = x(t_N)\}$. Then the mapping (14) can be rewritten as,

$$Y = f(x(\boldsymbol{t})) = \mathcal{F}(\boldsymbol{t}) = \mathcal{F}(t_1, t_2, \ldots, t_N).$$

The mean value is the most important statistical property of the random variable Y. In addition, if the process $x(t)$ is K-th order stationary, the dimension of the mean mapping for N observations is reduced by one.

Lemma 10. *The mean of $Y = f(x(\boldsymbol{t}))$ for N discrete time observations of a K-th order stationary random process has dimension $(N-1)$, i.e., if $N \leq K$,*

$$\bar{Y} = \mathrm{E}[f(x(\boldsymbol{t})] = \mathcal{C}(t_2 - t_1, t_3 - t_1, \ldots, t_N - t_1) = \mathcal{C}(\tau_1, \tau_2, \ldots, \tau_{N-1}) = \mathcal{C}(\boldsymbol{\tau})$$

where $\tau_i = t_{i+1} - t_1$, $i = 1, 2, \ldots, N-1$.

Proof. Assuming Lemma 1 and the first sample $X_1 = x(t_1)$ to be a reference, the joint probability density of N process observations becomes, $f_X(\boldsymbol{x}; 0, t_2 - t_1, \ldots, t_N - t_1) \equiv \tilde{f}_X(\boldsymbol{x}; t_2 - t_1, \ldots, t_N - t_1)$, so the corresponding statistical moments have dimension $(N-1)$. $\square$

In optimization problems, it is useful to consider the gradient of $\bar{Y}$, i.e.,

$$\nabla \bar{Y} = \left(\frac{\partial \bar{Y}}{\partial \tau_1}, \frac{\partial \bar{Y}}{\partial \tau_2}, \ldots, \frac{\partial \bar{Y}}{\partial \tau_{N-1}} \right) = \frac{\partial \bar{Y}}{\partial \mathcal{T}} \left(\frac{\partial \mathcal{T}}{\partial \tau_1}, \frac{\partial \mathcal{T}}{\partial \tau_2}, \ldots, \frac{\partial \mathcal{T}}{\partial \tau_{N-1}} \right)$$

where $\mathcal{T}$ is a application dependent measure of the vector $\boldsymbol{\tau}$ such as the norms, $\mathcal{T} = \|\boldsymbol{\tau}\|_1 = |\tau_1| + \cdots + |\tau_{N-1}|$, or $\mathcal{T} = \|\boldsymbol{\tau}\|_\infty = \max(\tau_1, \ldots, \tau_{N-1})$.

In general, assuming the Taylor's series defined in Theorem 1 on p. 8, a multivariate function of a random vector $\boldsymbol{x}$ can be expanded about the mean $\bar{\boldsymbol{x}} = \mathrm{E}[\boldsymbol{x}]$ as,

$$f(\bar{\boldsymbol{x}} + \boldsymbol{h}) \approx f(\bar{\boldsymbol{x}}) + \sum_{l=1}^{m} \sum_{|\boldsymbol{n}|_1 = l} \frac{\partial^{\boldsymbol{n}} f(\boldsymbol{x})}{\boldsymbol{n}!} \boldsymbol{h}^{\boldsymbol{n}}. \tag{15}$$

Thus, the value of $f(\bar{\boldsymbol{x}} + \boldsymbol{h})$ is a weighted sum of $\boldsymbol{h}^{\boldsymbol{n}}$ plus an offset $f(\bar{\boldsymbol{x}})$. More importantly, if the partial derivatives $\partial^{\boldsymbol{n}} f(\boldsymbol{x})$ are replaced with the coefficients $(l! \, p_l)$ which are independent of $\boldsymbol{n}$, the number of parameters in (15) is greatly reduced. Moreover, instead of precisely determining the values of $p_l \in \mathcal{R}$ to obtain the best possible approximation of the original function f, it is useful as well as sufficient to constrain the Taylor expansion (15) to the class of functions that are exactly constructed as,

$$f(\bar{\boldsymbol{x}} + \boldsymbol{h}) = f(\bar{\boldsymbol{x}}) + \sum_{l=1}^{m} p_l \sum_{|\boldsymbol{n}|_1 = l} \frac{l!}{\boldsymbol{n}!} \boldsymbol{h}^{\boldsymbol{n}} = \sum_{l=0}^{m} p_l (h_1 + \cdots + h_N)^l \tag{16}$$

where $p_0 = f(\bar{\boldsymbol{x}})$. The function expansion (16) represents a m-th degree polynomial in variable $|\boldsymbol{h}|_1$. The coefficients p_l of this polynomial can be set using Lemma 6, so the polynomial is convex.

The key realization is that the polynomial functions (16) have all the desired properties specified at the beginning of this section.

Claim 1. *A multivariate function $f : \mathcal{R}^N \mapsto \mathcal{R}$ having the desirable properties to measure the statistics of a random vector $\boldsymbol{x}$ is the m-th degree polynomial,*

$$Y = f(\boldsymbol{x}) = \sum_{l=0}^{m} p_l \left| \sum_{i=1}^{N} (X_i - \mathrm{E}[X_i]) \right|^l \tag{17}$$

where $p_0 = \mathrm{E}[\boldsymbol{x}]$, and the values of m and of the coefficients $p_l \in \mathcal{R}$ are determined by the application requirements.

The polynomial Function (17) can be used for any number of observations N. It is symmetric, so all observations are treated equally. Moreover, it is prone to integration and employing the expectation operator. The convexity can be achieved by Lemma 6.

4.1. Related Concepts

Assume the scalar function f defined in (17) for a K-th order stationarity random process $x(t)$. Define the auxiliary random variable,

$$Z(a) = \frac{1}{a} \sum_{i=1}^{N} (X_i - \bar{X}_i)$$

where a is a normalization constant, $a \neq 0$. The expression (17) can be then rewritten as, $Y(a) = \sum_{l=0}^{m} p_l Z^l(a)$. For $a = 1$, $Z(1)$ has a zero mean, and the variance, $\mathrm{var}[Z(1)] = \sum_{i,j} \mathrm{E}\left[(X_i - \bar{X}_i)(X_j - \bar{X}_j)\right]$. For $a = N$, $Z(N)$ represents a sample mean, and its variance is equal to, $\mathrm{var}[Z(N)] = \mathrm{var}[Z(1)]/N^2$. For $a = \sqrt{N}$, the variance of $Z(\sqrt{N})$ is normalized by the number of dimensions, i.e., $\mathrm{var}\left[Z(\sqrt{N})\right] = \mathrm{var}[Z(1)]/N$.

For $p_l = s^l/l!$, in the limit of large m, Equation (17) gives,

$$Y = \lim_{m \to \infty} \sum_{l=0}^{m} \frac{s^l}{l!} (Z(a))^l = \mathrm{e}^{s\,Z(a)}.$$

The mean, $\mathrm{E}[Y] = \mathrm{E}\left[\mathrm{e}^{s\,Z(a)}\right]$, is the moment generating function of the random variable $Z(a)$.

In data processing, the sample mean is intended to be an estimate of the true population mean, i.e., $N \gg 1$ is required. Here, $Z(a)$ is calculated over a finite number of vector or process dimensions, so it is a random variable for $\forall a \in \mathcal{R} \setminus \{0\}$. The variable $Z(N)$ should be then referred to as an arithmetic average or a center of gravity of the random vector $\boldsymbol{X}$ in the Euclidean space $\mathcal{R}^N$, i.e.,

$$Z(N) \triangleq \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \in \mathcal{R}. \tag{18}$$

Note that (18) is not an l_1-norm, since the variables X_i are not summed with absolute values.

If the random variables X_i are independent, the distribution of $Z(a)$ is given by convolution of their marginal distributions. For correlated observations, if the characteristic function, $\tilde{f}(s/a) = \mathrm{E}\left[\mathrm{e}^{\mathrm{j}\cdot|X|_1 s/a}\right]$, of the sum $|\boldsymbol{X}|_1$ can be obtained, the distribution of $Z(a) = |\boldsymbol{X}|_1/a$ is computed as the inverse transform,

$$f_Z(Z(a)) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{j}sZ(a)} \tilde{f}\left(\frac{s}{a}\right) \mathrm{d}s.$$

Many other properties involving the sums of random variables can be obtained. For instance, if the random variables X_i are independent and have zero mean, then,

$$\mathrm{E}[Z^m(1)] - \mathrm{E}[(Z(1) - X_N)^m] = \begin{cases} \mathrm{E}\left[X_N^m\right] & m = 2,3 \\ \mathrm{E}\left[X_N^4\right] + 6\sum_{i=1}^{N-1} \mathrm{E}\left[X_N^2\right]\mathrm{E}\left[X_i^2\right] & m = 4 \\ \mathrm{E}\left[X_N^5\right] + 10\sum_{i=1}^{N-1} \mathrm{E}\left[X_N^2\right]\mathrm{E}\left[X_i^3\right] + \mathrm{E}\left[X_N^3\right]\mathrm{E}\left[X_i^2\right] & m = 5. \end{cases}$$

Considering Claim 1, an important statistic for a random vector can be defined as the m-th central sum-moment.

Definition 4. *The m-th central sum-moment of random vector $\boldsymbol{X} \in \mathcal{R}^N$ is computed as,*

$$\Sigma\mu_m(\boldsymbol{X}) = \mu_m(|\boldsymbol{X}|_1) = \mathrm{E}\left[\left|\sum_{i=1}^{N}(X_i - \bar{X}_i)\right|^m\right], \quad m = 1, 2, \dots$$

Lemma 11. *The second central sum-moment of random vector is equal to the sum of all elements of its covariance matrix, i.e.,*

$$\Sigma\mu_2(\boldsymbol{X}) = \mu_2(|\boldsymbol{X} - \bar{\boldsymbol{X}}|_1) = \mathrm{E}\left[\left(\sum_{i=1}^{N}(X_i - \bar{X}_i)\right)^2\right] = \sum_{i,j=1}^{N} \mathrm{cov}\left[X_i, X_j\right] = |\boldsymbol{C}_x|_1.$$

Furthermore, the second central sum-moment is also equal to the variance of $|\boldsymbol{X}|$, i.e.,

$$\Sigma\mu_2(\boldsymbol{X}) = \mathrm{var}[\,|\boldsymbol{X}|_1\,] = \mathrm{var}\left[\sum_{i=1}^{N}(X_i - \bar{X}_i)\right].$$

Proof. The first equality is shown by expanding the expectation, and substituting for elements of the covariance matrix, $[\boldsymbol{C}_x]_{i,j} = \mathrm{cov}\left[X_i, X_j\right]$. The second expression follows by noting that $\sum_{i=1}^{N}(X_i - \bar{X}_i)$ has zero mean. □

In the literature, there are other measures involving sums of random variables. For instance, in Mean-Field Theory, the model dimensionality is reduced by representing N-dimensional vectors by their center of gravity [20]. The central point of a vector is also used in the first order approximation of multivariate functions in [21] and in the model overall variance in [22].

In Measure Theory [23], the total variation (TV) of a real-valued univariate function, $x : (t_0, t_N) \mapsto \mathcal{R}$, is defined as the supremum over all possible partitions $\mathcal{P} : t_0 < t_1 < \cdots < t_N$ of the interval (t_0, t_N), i.e.,

$$\mathrm{TV}(x) = \sup_{\mathcal{P}} \sum_{i=0}^{N-1} |x(t_{i+1} - x(t_i)|.$$

The TV concept can be adopted for observations $X_i = x(t_i)$ of a stationary random process $x(t)$ at discrete time instances, $\{t_i\}_{0:N}$. A mathematically tractable mean TV measure can be defined as,

$$\overline{\mathrm{TV}}^2(x) = \sup_{\mathcal{P}} \mathrm{E}\left[\sum_{i=0}^{N-1} |X_{i+1} - X_i|^2\right] = \sup_{\mathcal{P}} 2N(\mathrm{E}\left[X_i^2\right] - \mathrm{cov}[X_{i+1}, X_i]).$$

Jensen's inequality for a random vector assuming equal weights can be stated as [17],

$$\mathrm{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}_i)\right|^m\right] \leq \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}[|X_i - \bar{X}_i|^m].$$

Alternatively, exchanging the expectation and summation, Jensen's inequality becomes,

$$\sum_{i=1}^{N} \left| E[X_i - \bar{X}_i] \right|^m \le E \left[\sum_{i=1}^{N} |X_i - \bar{X}_i|^m \right]. \tag{19}$$

Furthermore, if the right-hand side of (19) is to be minimized and $m = 2$, the inequality in (19) changes to equality. In particular, consider the minimum mean square error (MMSE) estimation of a vector of random parameters $\boldsymbol{P} = \{P_i\}_{1:N}$ from measurements $\boldsymbol{X}$. Denoting $\bar{P}_i(\boldsymbol{X}) = E[P_i|\boldsymbol{X}]$, conditioned on $\boldsymbol{X}$, the MMSE estimator $\hat{\boldsymbol{P}}(\boldsymbol{X})$ minimizes [15],

$$
\begin{aligned}
\min_{\hat{P}|\boldsymbol{X}} E \left[\sum_{i=1}^{N} (\hat{P}_i(\boldsymbol{X}) - P_i)^2 \right] &= \min_{\hat{P}|\boldsymbol{X}} E \left[\sum_{i=1}^{N} \left((\hat{P}_i(\boldsymbol{X}) - \bar{P}_i(\boldsymbol{X})) - (P_i - \bar{P}_i(\boldsymbol{X})) \right)^2 \right] \\
&= \min_{\hat{P}|\boldsymbol{X}} \left\{ \sum_{i=1}^{N} (\hat{P}_i(\boldsymbol{X}) - \bar{P}_i(\boldsymbol{X}))^2 + E \left[\sum_{i=1}^{n} (P_i - \bar{P}_i(\boldsymbol{X}))^2 \right] \right\} \\
&= \min_{\hat{P}|\boldsymbol{X}} \sum_{i=1}^{N} (E[\hat{P}_i(\boldsymbol{X}) - P_i|\boldsymbol{X}])^2 = \min_{\hat{P}|\boldsymbol{X}} \sum_{i=1}^{N} (\hat{P}_i(\boldsymbol{X}) - E[P_i|\boldsymbol{X}])^2
\end{aligned}
\tag{20}
$$

where the expectations are over the conditional distribution $f_{\boldsymbol{P}|\boldsymbol{X}}$.

In signal processing, a length N moving average (MA) filter transforms the input sequence X_i into an output sequence Y_i by discrete-time convolution $\circledast$, i.e.,

$$Y_i = \sum_{j=0}^{N-1} X_{i-j} = [\underbrace{1\,1\,\ldots\,1}_{\mathbf{1}_N}] \circledast X_i.$$

The (auto-) correlations of the input and output sequences are related by Lemma 3 on p. 6, i.e.,

$$C_Y(i) = \mathbf{1}_N \circledast \mathbf{1}_N \circledast C_x(i) = \sum_{j=-N+1}^{N-1} (N - |j|)\, C_x(i-j). \tag{21}$$

Note that if the input process is stationary, then the input and output processes are jointly stationary.

4.2. Multiple Random Processes

The major complication with observing, evaluating, and processing multiple random processes is how to achieve their time alignment and amplitude normalization (scaling). Focusing here on the time alignment problem only, denote the discrete time observation instances of L random processes as,

$$\boldsymbol{t}_l = \{t_{l1} < t_{l2} < \ldots < t_{lN_l}\}, \quad l = 1, 2, \ldots, L.$$

Assume that the first time instance t_{l1} of every process serves as a reference. Then, there are $(L-1)$ uncertainties in time alignment of L processes, i.e.,

$$\Delta_l = (t_{l1} - t_{11}) \in \mathcal{R}, \quad l = 2, 3, \ldots, L.$$

The $(L-1)$ values Δ_l are unknown parameters which must be estimated. Note also that the difference,

$$\Delta_l - \Delta_k = t_{l1} - t_{k1}$$

represents an unknown time shift between the process $x_l(t)$ and $x_k(t)$.

For any multivariate stationary distribution of observations of two random processes, the corresponding cross-correlation normally attains a maximum for some time shift between these processes [12]. Hence, a usual strategy for aligning the observed sequences

is to locate the maximum value of their cross-correlation. The time shifts Δ_l, $l = 1, 2, \ldots, L$, are then estimated as,

$$\hat{\Delta}_l = \underset{\Delta}{\text{argmax}}\, C_{x_1 x_l}(\Delta), \quad \Delta \in \{(t_{li} - t_{11})\}_{i=1,\ldots,N_l}.$$

Assuming the center values $\bar{X}_1$ and $\bar{X}_2$ in (18) as scalar representations of the vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, their cross-covariance can be computed as,

$$\text{cov}[|\mathbf{X}_1|_1, |\mathbf{X}_2|_1] = N^2 \text{cov}[\bar{X}_1, \bar{X}_2] = N^2\, \text{E}[(\bar{X}_1 - \text{E}[X_1])(\bar{X}_2 - \text{E}[X_2])]. \tag{22}$$

The task now is how to generalize the pairwise cross-covariance (22) to the case of multiple random vectors having possibly different lengths. If all random vectors of interest are concatenated into one single vector, the *m*-th joint central sum-moment can be defined by utilizing Claim 1.

Definition 5. *The m-th central sum-moment for L random processes with N_l observations, $l = 1, 2, \ldots, L$, is computed as,*

$$\Sigma \mu_m(\mathbf{X}_1, \ldots, \mathbf{X}_L) = \mu_m(|\mathbf{X}_1|_1 + \ldots + |\mathbf{X}_L|_1) = \text{E}\left[\left|\sum_{l=1}^{L}\sum_{i=1}^{N_l}(X_{li} - \bar{X}_{li})\right|^m\right].$$

Lemma 12. *The second central sum-moment for L random processes with N_l observations, $l = 1, 2, \ldots, L$, is equal to the sum of all pairwise covariances, i.e.,*

$$\Sigma \mu_2(\mathbf{X}_1, \ldots, \mathbf{X}_L) = \sum_{l,k=1}^{L}\sum_{i,j=1}^{N_l} \text{cov}\left[X_{li}, X_{kj}\right] = \sum_{l,k=1}^{L} |\text{cov}[\mathbf{X}_l, \mathbf{X}_k]|_1 = \text{var}[|\mathbf{X}_1|_1 + \ldots + |\mathbf{X}_L|_1].$$

Proof. The expression can be obtained by expanding the sum and then applying the expectation. □

Many other properties of central and noncentral sum-moments can be obtained. For example, assuming two equal-length vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, it is straightforward to show that,

$$\text{E}\left[(|\mathbf{X}_1|_1 + |\mathbf{X}_2|_1)^2\right] - \text{E}\left[|\mathbf{X}_1|_1^2 + |\mathbf{X}_2|_1^2\right] = 2\sum_{i,j=1}^{N} \text{E}[X_{1i}X_{2j}]$$

$$\text{E}\left[(|\mathbf{X}_1|_1 + |\mathbf{X}_2|_1)^2\right] - \text{E}\left[\|\mathbf{X}_1 - \mathbf{X}_2\|_2^2\right] = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \text{E}[X_{1i}X_{1j} + X_{2i}X_{2j}] + 2\sum_{i,j=1}^{N} \text{E}[X_{1i}X_{2j}] + 2\sum_{i=1}^{N} \text{E}[X_{1i}X_{2i}].$$

5. Illustrative Examples

This section provides examples to quantitatively evaluate the results developed in the previous sections. In particular, the accuracy of approximate linear LS regression proposed in Section 3.1 is assessed to justify its lower computational complexity. The central sum-moments introduced in Section 4 are compared assuming correlated Gaussian processes. Finally, several signal processing problems involving the 1st order Markov process are investigated.

5.1. Linear Regression

Consider a classical one-dimensional linear LS regression problem with independent and identically normally distributed errors. The errors are also assumed to be independent from all other data model parameters. The data points are generated as,

$$X_i = \Delta i : Y_i = P_2 X_i + P_1 + E_i, \quad i = 1, 2, \ldots, N$$

where E_i are zero-mean, uncorrelated Gaussian samples having the equal variance σ_e^2, and P_1 and P_2 are unknown parameters to be estimated. This LS problem can be solved exactly using the expression (2), and substituting $w_{1i} = 1$ and $w_{2i} = \Delta i$, $\forall i = 1, 2, \dots, N$. Alternatively, to avoid inverting the $N \times N$ data matrix, the procedure devised in Section 3.1 suggests to split the data into two equal-size subsets, compute the average data point for each subset, and then solve the corresponding set of two equations with two unknowns. Specifically, the set of two equations with the unknown parameters $\hat{P}_1$ and $\hat{P}_2$ is,

$$
\begin{aligned}
\frac{2}{N} \sum_{i=1}^{N/2} Y_i &= \hat{P}_1 + \hat{P}_2 \frac{\Delta}{8} N(2+N) \\
\frac{2}{N} \sum_{i=N/2+1}^{N} Y_i &= \hat{P}_1 + \hat{P}_2 \frac{\Delta}{8} N(2+3N)
\end{aligned}
\tag{23}
$$

assuming N is even, and using, $\sum_{i=1}^{N/2} \Delta i = \frac{\Delta}{8} N(2+N)$, and $\sum_{i=N/2+1}^{N} \Delta i = \frac{\Delta}{8} N(2+3N)$. Denoting $\bar{Y}_1 = \frac{2}{N} \sum_{i=1}^{N/2} Y_i$ and $\bar{Y}_2 = \frac{2}{N} \sum_{i=N/2+1}^{N} Y_i$, the closed-form solution of (23) is,

$$
\begin{aligned}
\hat{P}_1 &= \frac{1}{2N}((2+3N)\bar{Y}_1 - (2+N)\bar{Y}_2) \\
\hat{P}_2 &= \frac{4}{\Delta N^2}(\bar{Y}_2 - \bar{Y}_1).
\end{aligned}
$$

As a numerical example, assume the true values $P_1 = 1.5$, $P_2 = 0.3$, $\mathrm{E}[E_i] = 0$, $\mathrm{var}[E_i] = 1$, and $N = 40$ data points. Figure 2 shows the intervals $(\bar{T} - \sqrt{\mathrm{var}[T]}, \bar{T} + \sqrt{\mathrm{var}[T]})$ versus the subset size $1 \leq N_1 \leq N/2$ for the random variable T defined as,

$$
T = 100 \, \frac{S_{\mathrm{apr}}(N) - S_{\mathrm{opt}}(N)}{S_{\mathrm{opt}}(N)}
$$

where $S_{\mathrm{apr}}(N) = \sum_{i=1}^{N}(Y_i - \hat{P}_1 - \hat{P}_2 X_i)^2$ and $S_{\mathrm{opt}}(N) = \sum_{i=1}^{N}(Y_i - P_1 - P_2 X_i)^2$ are the total MSEs. In the limit, $\lim_{N \to \infty}(S_{\mathrm{apr}}(N) - S_{\mathrm{opt}}(N)) = 0$, since a sufficiently large subset of data is as good as the complete set of data. For finite N, it is likely that $S_{\mathrm{apr}}(N) > S_{\mathrm{opt}}(N)$, so the lower bounds in Figure 2 converge much faster to zero than the upper bounds.
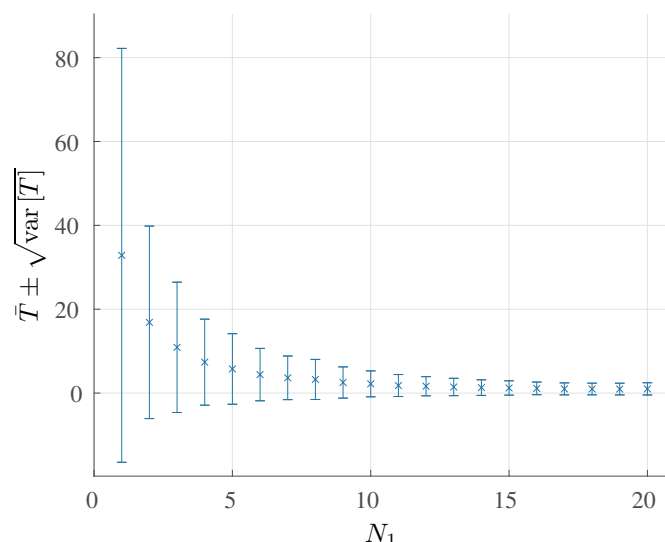

Figure 2. The relative total mean-square error of the approximate linear LS regression.

5.2. Comparison of Central Moments

Assuming Lemma 11 and Equation (3), the second central sum-moment of the 2nd order Markov process of length N is,

$$
\Sigma \mu_2(\mathbf{X}) = \sum_{i,j=1}^{N} C_{2\mathrm{MP}}(i-j) = \sigma^2 \left(N + 2 \sum_{i=1}^{N-1} i(1 + (N-i)\alpha)\, e^{-\alpha(N-i)} \right).
$$

These moments are compared in Figure 3 for different values of the sequence length N, three values of the parameter α, and assuming $\sigma^2 = 1$. It can be observed that the values of the second central sum-moment are increasing with N and the level of correlation, $e^{-\alpha}$.
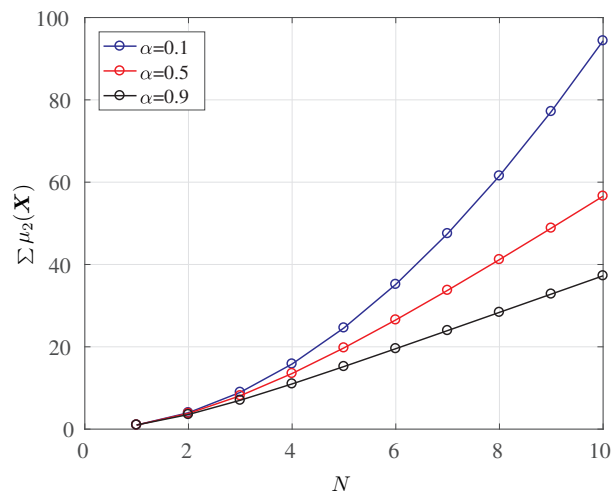

Figure 3. The second central sum-moments of the 2nd order Markov process with parameter α and length N.

Consider now the following three central moments of order $m = 1, 2, \ldots,$ i.e.,

$$\bar{S}_{\mathrm{mnk}}(N) = \sum_{i=1}^{N} \mathrm{E}\left[|\sqrt{N}\,X_i|^m\right]$$

$$\Sigma\mu_2(N) = \mathrm{E}\left[\left|\sum_{i=1}^{N} X_i\right|^m\right]$$

$$\Sigma\tilde{\mu}_2(N) = \mathrm{E}\left[\left(\sum_{i=1}^{N} |X_i|\right)^m\right].$$

The moment, $\bar{S}_{\mathrm{mnk}}$, is the mean Minkowski distance; the scaling by $\sqrt{N}$ is introduced to facilitate the comparison with the other two moments, i.e., the mean sum-moment $\Sigma\mu_2$, and the mean sum-moment $\Sigma\tilde{\mu}_2$ having the samples summed as the l_1 norm. More importantly, assuming a correlated Gaussian process $X_i = x(t_i)$, the central sum-moment can be readily obtained in a closed form whereas obtaining the closed form expressions for the other two metrics may be mathematically intractable. In particular, by factoring the covariance matrix as, $C_x = T_x T_x^T$, the correlated Gaussian vector can be expressed as, $X = TU$. Then the sum of elements, $|X|_1 = \mathbf{1}^T TU$, and the m-th central sum-moment can be computed as,

$$\Sigma\mu_2(N) = \mathrm{E}\left[\left|\mathbf{1}^T TU\right|^m\right] = \left\|\mathbf{1}^T T\right\|_2^m \mathrm{E}[|U|^m] = \left\|\mathbf{1}^T T\right\|_2^m \frac{2^{m/2}}{\sqrt{\pi}} \Gamma\left(\frac{m+1}{2}\right)$$

where U is a zero-mean, unit-variance Gaussian random variable, and Γ denotes the gamma function.

Figure 4 shows all three moments as a function of sequence length N for three values of parameter α assuming the 1st order Markov process. The vertical axis in Figure 4 is scaled by $1/N$ for convenience. Note that, for uncorrelated, i.e., independent Gaussian samples, the moments $\Sigma\mu_2$ and $\Sigma\tilde{\mu}_2$ are identical. More importantly, all three moments are strictly increasing with the number of samples N and with the moment order m.
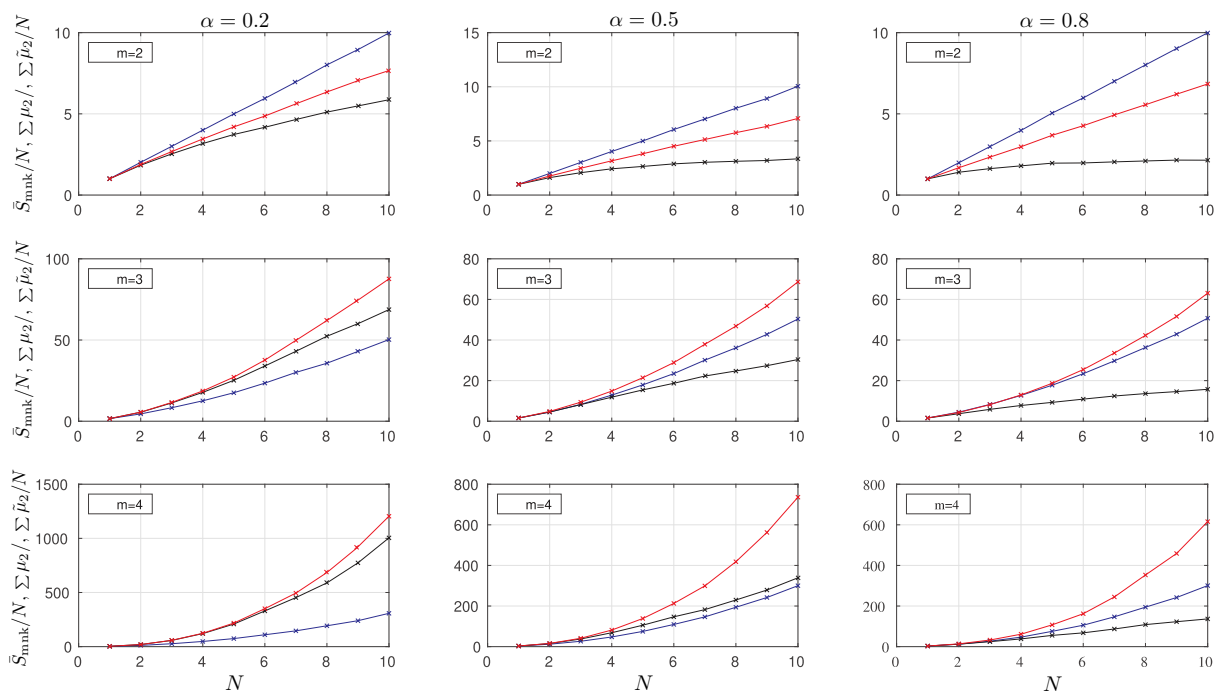
Figure 4. The Minkowski (blue), sum-moment (black), and sum-moment of absolute values (red) mean statistics for the 1st order Markov sequence of length N. Columns: different values α. Rows: different values m.

Finally, the second central sum-moment can be visualized in two dimensions. Consider $N = 3$ observations of a zero-mean real-valued stationary random process, $X_i = x(t_i)$, $i = 1, 2, 3$. Let $\sum \mu_2 = \mathrm{E}\big[(X_1 + X_2 + X_3)^2\big] = \sum_{i,j=1}^3 \mathrm{E}\big[X_i X_j\big]$. Assuming the 1st order Markov correlation model, $\mathrm{E}\big[X_i X_j\big] \propto e^{-0.5|\tau|}$, Figure 5 shows the values of the central sum-moment $\sum \mu_2$ versus the sample distances $\tau_1 = (t_2 - t_1)$ and $\tau_2 = (t_3 - t_1)$. Several symmetries can be observed in Figure 5. In particular, the central sum-moment $\sum \mu_2$ is symmetric about the axis $\tau_1 = \tau_2$ as well as about the axis $\tau_1 = -\tau_2$. These symmetries are consequences of the following equalities:

$$\left.\begin{array}{rcl} \sum \mu_2(\tau_1, \tau_2) & = & \sum \mu_2(\tau_2, \tau_1) \\ \sum \mu_2(\tau_1, \tau_2) & = & \sum \mu_2(-\tau_1, -\tau_2) \end{array}\right\} \forall \tau_1, \tau_2.$$
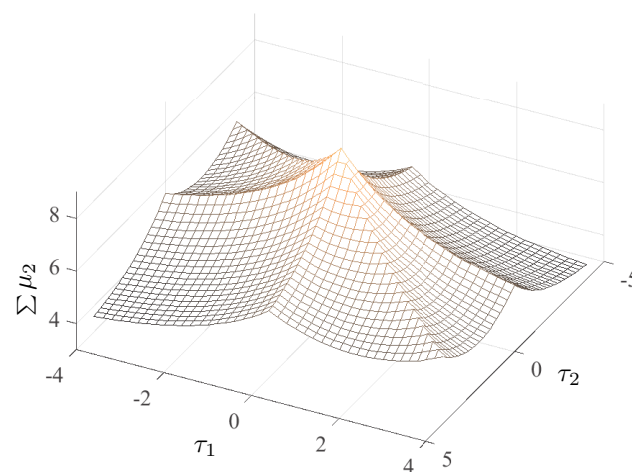

Figure 5. The second central sum-moment as a function of time differences between $N = 3$ observations of a stationary random process.

5.3. Signal Processing Problems for the 1st Order Markov Process

Consider the 1st order Markov process observed at the output of a length N MA filter. According to (21), the (auto-) covariance of the output process is,

$$C_{\text{1MP+MA}}(k; \alpha) = \sum_{j=-N+1}^{N-1} (N - |j|)\sigma^2 \, e^{-\alpha|k-j|}. \tag{24}$$

Conjecture 1. *The MA filtering of the 1st order Markov process generates nearly a 2nd order Markov process.*

The parameter of the 2nd order Markov process approximating the combined (auto-) covariance (24) can be obtained using the LS regression fit, i.e.,

$$\hat{\alpha} = \underset{\tilde{\alpha}}{\text{argmin}} \sum_k \left(C_{\text{1MP+MA}}(k; \alpha) - C_{\text{2MP}}(k; \tilde{\alpha}) \right)^2. \tag{25}$$

Substituting (3) and (24) into (25), and letting the first derivative to be equal to zero, the LS estimate $\hat{\alpha}$ must satisfy the linear equation,

$$\sum_k \hat{\alpha}|k| + 1 + W_{-1}\left(-\frac{C_{\text{1MP+MA}}(k; \alpha)}{e} \right) = 0$$

which can be readily solved for $\hat{\alpha}$, and W_{-1} denotes the Lambert function [24].

A discrete time sequence of N elements has the (auto-) covariance constrained to $(2N - 1)$ time indexes as indicated in (24). Assuming the length N MA filter, and that there are $(n_x N)$ samples, $n_x = 1, 2, \ldots,$ of the 1st order Markov process available, the (auto-) covariance (24) has the overall length $2N(n_x + 1) - 3$ samples. Figure 6 compares the MSE,

$$\text{MSE}(n_x) = 100 \times \frac{\sum_k (C_{\text{1MP+MA}}(k; \alpha) - C_{\text{2MP}}(k; \hat{\alpha}))^2}{\sum_k C_{\text{1MP+MA}}^2(k; \alpha)}$$

of the LS fit of the (auto-) covariance of the 2nd order Markov process to the combined (auto-) covariance of the 1st order Markov process and the impulse response of the MA filter assuming three values of α and two values of n_x. Given α and n_x, Figure 6 shows that the best LS fit occurs for a certain value of the MA filter length N. It can be concluded that, in general, the 1st order Markov process changes to the 2nd order Markov process at the output of the MA filter.
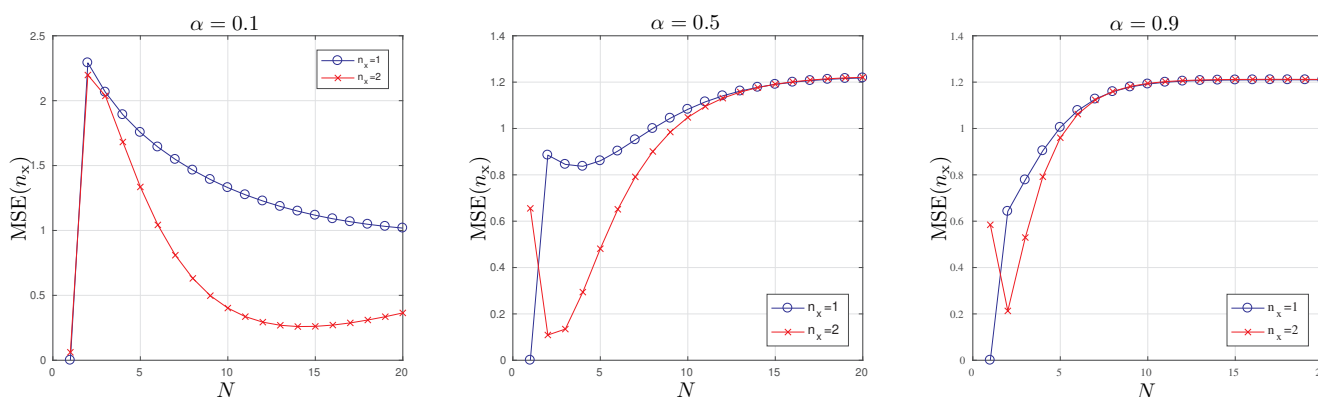

Figure 6. The MSE of approximating the (auto-) covariance of the 1st order Markov process at the output of length N MA filter by the (auto-) covariance of the 2nd order Markov process.

The second problem to investigate is a linear MMSE (LMMSE) prediction of the 1st order Markov process observed at the output of a MA filter. In particular, given N samples

$X_i = x(t_i)$, $i = 1, 2, \ldots, N$, of random process having the (auto-) covariance (24), the task is to predict its future value, $X_{N+1} = x(t_{N+1})$, $t_{N+1} > t_N$.

In general, the impulse response h of the LMMSE filter to estimate an unknown scalar parameter P from the measurements X is computed as [15],

$$h = \mathrm{E}\left[(x - \bar{x})(X - \bar{X})^T\right]\mathrm{E}[P\,X]. \tag{26}$$

Here, the unknown parameter $P = X_{N+1}$, and $\mathrm{E}[X_{N+1}X_i] = C_{\mathrm{1MP+MA}}(N + 1 - i)$ and $\mathrm{E}[X_iX_j] = C_{\mathrm{1MP+MA}}(i - j)$ in (26) which gives the length N LMMSE filter,

$$h = [\underbrace{0\;0\;\ldots\;0}_{N-1}\; C_{\mathrm{1MP+MA}}(1)].$$

Consequently, the predicted value, $X_{N+1} = X_N\, C_{\mathrm{1MP+MA}}(1)/\sigma_X^2$. Note that the same procedure, but excluding the MA filter, gives the LMMSE estimate, $X_{N+1} = X_N\, C_{\mathrm{1MP}}(1)/\sigma_X^2$.

The last problem to consider is a time alignment of two zero-mean, jointly stationary processes. It is assumed that the normalized cross-covariance of these two processes is,

$$\frac{\mathrm{E}[X_{1i}X_{2j}]}{\sqrt{\mathrm{E}[X_{1i}^2]\mathrm{E}[X_{2i}^2]}} = \mathrm{e}^{-\alpha|i-j|}(1 + \alpha|i - j|). \tag{27}$$

Denote the uncertainty in determining the difference, $(i - j)$, as Δ. In order to estimate the unknown parameters α and Δ, the left-hand side of (27) can be estimated by the method of moments, i.e., let $\mathrm{E}[X_i^2] \approx \frac{1}{N}\sum_{i=1}^{N} X_i^2$, and, $\mathrm{E}[X_{1i}X_{2j}] \approx \frac{1}{N}\sum_{i=1}^{N} X_{1i}X_{2|i-\Delta|}$. The cross-covariance (27) can be then rewritten as,

$$v_k = \frac{\sum_{i=1}^{N} X_{1i}X_{2|i-\Delta-k|}}{\sqrt{\sum_{i=1}^{N} X_{1i}^2}\sqrt{\sum_{i=1}^{N} X_{2i}^2}} = \mathrm{e}^{-\alpha|\Delta+k|}(1 + \alpha|\Delta + k|), \quad k = 0, 1, 2, \ldots$$

Utilizing the Lambert function W_{-1}, the cross-covariance can be rewritten further as,

$$\alpha|\Delta + k| = -1 - W_{-1}\left(-\frac{v_k}{\mathrm{e}}\right) \triangleq \tilde{v}_k, \quad k = 0, 1, 2, \ldots \tag{28}$$

Assuming, without loss of generality, that $\Delta \geq 0$, the absolute value in (28) can be ignored. Consequently, the unknown parameters α and Δ can be obtained as a linear LS fit to N measured and calculated values $\tilde{v}_k$ in the linear model (28).

6. Conclusions

The development of a novel statistical measure to enable correlation analysis for multiple random vectors resumed by summarizing background knowledge on statistical description of discrete time random processes. This was then extended with the derivation of several supporting results which were used in the following sections. Specifically, it was shown that linear regression can be effectively approximated by splitting the data into disjoint subsets and assuming only one average data point within each subset. In addition, a procedure for generating multiple Gaussian processes with the prescribed autocovariance and cross-covariance was devised. The main result of the paper was obtained by assuming the Taylor's expansion of multivariate symmetric scalar functions, and then approximating the Taylor's expansion by a univariate polynomial. The single polynomial variable is a simple sum of variables in the original multivariate function. The polynomial approximation represents a mapping from multiple discrete time observations of a random process to a multidimensional scalar field. The mean field value is a weighted sum of canonical central moments with increasing orders. These moments were named central sum-moments to reflect how they are defined. The sum-moments were then discussed in light of other similar concepts such as total variance, Mean Field Theory, and moving average sequence

filtering. Illustrative examples were studied in the last section of the paper. In particular, the accuracy of approximate linear regression was evaluated quantitatively assuming two disjoint data subsets. Assuming the 1st and the 2nd order Markov processes, the central sum-moments were compared with the mean Minkowski distance. For Gaussian processes, the central sum-moments can be obtained in closed form. The remaining problems investigated moving average filtering of the 1st order Markov processes and its prediction using a linear MMSE filter.

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.

Abbreviations

The following abbreviations are used in the paper:

1MP	1st order Markov process
2MP	2nd order Markov process
AR	autoregressive
LMMSE	linear minimum mean square error
LS	least squares
MMSE	minimum mean square error
MSE	mean square error
MA	moving average
TV	total variance
$|\cdot|$	absolute value, set cardinality, sum of vector elements
$\mathrm{E}[\cdot]$	expectation
$\mathrm{corr}[\cdot]$	correlation
$\mathrm{cov}[\cdot]$	covariance
$\mathrm{var}[\cdot]$	variance
$(\cdot)^{-1}$	matrix inverse
$(\cdot)^{T}$	matrix/vector transpose
f_x	distribution of a random variable X
$f, \dot{f}, \ddot{f}$	function f, and its first and second derivatives
$\mathbb{N}_+$	positive non-zero integers
$\mathcal{R}, \mathcal{R}^+$	real numbers, positive real numbers
$\bar{X}$	mean value of random variable X
X_{ij}	j-th sample of process i
W_{-1}	Lambert function

References

1. Drezner, Z. Multirelation—A correlation among more than two variables. *Comput. Stat. Data Anal.* **1995**, *19*, 283–292. [CrossRef]
2. Dear, R.; Drezner, Z. On the significance level of the multirelation coefficient. *J. Appl. Math. Decis. Sci.* **1997**, *1*, 119–130. [CrossRef]
3. Geiß, S.; Einax, J. Multivariate correlation analysis—A method for the analysis of multidimensional time series in environmental studies. *Chemom. Intell. Lab. Syst.* **1996**, *32*, 57–65. [CrossRef]
4. Abdi, H. Chapter Multiple correlation coefficient. In *Encyclopedia of Measurements and Statistics*; SAGE: Los Angeles, CA, USA, 2007; pp. 648–655. [CrossRef]
5. Székely, G.J.; Rizzo, M.L.; Bakirov, N.K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **2007**, *35*, 2769–2794. [CrossRef]
6. Merigó, J.M.; Casanovas, M. A New Minkowski Distance Based on Induced Aggregation Operators. *Int. J. Comput. Intell. Syst.* **2011**, *4*, 123–133. [CrossRef]
7. Nguyen, H.V.; Müller, E.; Vreeken, J.; Efros, P.; Böhm, K. Multivariate maximal correlation analysis. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. II–775–II–783. [CrossRef]
8. Josse, J. Measuring multivariate association and beyond. *Stat. Surv.* **2016**, *10*, 132–167. [CrossRef] [PubMed]

9. Shu, X.; Zhang, Q.; Shi, J.; Qi, Y. A comparative study on weighted central moment and its application in 2D shape retrieval. *Information* **2016**, *7*, 10. [CrossRef]

10. Böttcher, B.; Keller-Ressel, M.; Schilling, R.L. Distance multivariance: New dependence measures for random vectors. *Ann. Stat.* **2019**, *47*, 2757–2789. [CrossRef]

11. Wang, J.; Zheng, N. Measures of correlation for multiple variables. *arXiv* **2020**, arXiv:1401.4827.

12. Gardner, W.A. *Introduction to Random Processes With Applications*, 2nd ed.; McGraw-Hill: New York, NY, USA, 1990.

13. Papoulis, A.; Pillai, S.U. *Probability, Random Variables, and Stochastic Processes*, 4th ed.; McGraw-Hill: New York, NY, USA, 2002.

14. Giller, G.L. The Statistical Properties of Random Bitstreams and the Sampling Distribution of Cosine Similarity. *SSRN Preprint* **2012**, [CrossRef]

15. Kay, S.M. *Fundamentals of Statistical Signal Processing: Estimation Theory*; Prentice Hall: Upper Saddle River, NJ, USA, 1993; Volume I.

16. Oppenheim, A.V.; Schafer, R.W.; Buck, J.R. *Discrete-Time Signal Processing*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2009.

17. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.

18. Folland, G.B. Higher-Order Derivatives and Taylor's Formula in Several Variables. Available online: https://sites.math. washington.edu/~folland/Math425/taylor2.pdf (accessed on 1 December 2020).

19. Apostol, T. *Mathematical Analysis*, 2nd ed.; Pearson: London, UK, 1973.

20. Yedidia, J.S. chapter An idiosyncratic journey beyond mean field theory. In *Advanced Mean Field Methods*; MIT Press: Cambridge, MA, USA, 2001; pp. 21–36.

21. Sobol, I.M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul.* **2001**, *55*, 271–280. [CrossRef]

22. Saltelli, A.; Ratto, M.; Tarantola, S.; Campolongo, F. Sensitivity analysis for chemical models. *Chem. Rev.* **2005**, *105*, 2811–2828. [CrossRef] [PubMed]

23. Shirali, S. *A Concise Introduction to Measure Theory*; Springer: Berlin/Heidelberg, Germany, 2018.

24. Abramowitz, M.; Stegun, I.A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; Dover: Mineola, NY, USA, 1974.