Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2015-028 September 21, 2015

Network Maximal Correlation

Soheil Feizi, Ali Makhdoumi, Ken Duffy, Manolis Kellis, and Muriel Medard

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Network Maximal Correlation

Soheil Feizi[1,3], Ali Makhdoumi[1,3], Ken Duffy[2], Manolis Kellis[1] and Muriel Médard[1]

September 2015

Abstract

Identifying nonlinear relationships in large datasets is a daunting task particularly when the form of the nonlinearity is unknown. Here, we introduce Network Maximal Correlation (NMC) as a fundamental measure to capture nonlinear associations in networks without the knowledge of underlying nonlinearity shapes. NMC infers, possibly nonlinear, transformations of variables with zero means and unit variances by maximizing total nonlinear correlation over the underlying network. For the case of having two variables, NMC is equivalent to the standard Maximal Correlation. We characterize a solution of the NMC optimization using geometric properties of Hilbert spaces for both discrete and jointly Gaussian variables. For discrete random variables, we show that the NMC optimization is an instance of the Maximum Correlation Problem and provide necessary conditions for its global optimal solution. Moreover, we propose an efficient algorithm based on Alternating Conditional Expectation (ACE) which converges to a local NMC optimum. For this algorithm, we provide guidelines for choosing appropriate starting points to jump out of local maximizers. We also propose a distributed algorithm to compute a 1-ϵ approximation of the NMC value for large and dense graphs using graph partitioning. For jointly Gaussian variables, under some conditions, we show that the NMC optimization can be simplified to a Max-Cut problem, where we provide conditions under which an NMC solution can be computed exactly. Under some general conditions, we show that NMC can infer the underlying graphical model for functions of latent jointly Gaussian variables. These functions are unknown, bijective, and can be nonlinear. This result broadens the family of continuous distributions whose graphical models can be characterized efficiently. We illustrate the robustness of NMC in real world applications by showing its continuity with respect to small perturbations of joint distributions. We also show that sample NMC (NMC computed using empirical distributions) converges exponentially fast to the true NMC value. Finally, we apply NMC to different cancer datasets including breast, kidney and liver cancers, and show that NMC infers gene modules that are significantly associated with survival times of individuals while they are not detected using linear association measures.

1 Introduction

Identifying relationships among variables in large datasets is an increasingly important task in systems biology [1], social sciences [2], finance [3], etc. While correlation-based measures capture linear associations, they can fail to infer true nonlinear relationships among variables, which can often occur in real-world applications [4]. One family of measures to infer nonlinear associations among

[1] Massachusetts Institute of Technology (MIT), Cambridge, US.
[2] Hamilton Institute, Maynooth University, Ireland.
[3] These authors contributed equally to this work.

variables is based on mutual information [5, 6]. Although mutual information computes a measure of association strength among variables, it does not provide functions through which variables are related to each other. Moreover, reliable computation of mutual information requires an excessive number of samples, particularly for large number of variables [7].

A classical measure to infer a nonlinear relationship between two variables is *Maximal Correlation* (MC), introduced by Gebelein [8] and studied in references [9–12]. MC infers, possibly nonlinear, transformations of two variables with zero means and unit variances by maximizing their pairwise correlation. MC can be computed efficiently for both discrete [13] and continuous [14] random variables. For discrete variables, under some mild conditions, MC is equal to the second largest singular value of a normalized joint probability distribution matrix [13]. In that case, transformations of variables can be characterized using right and left singular vectors of the normalized probability distribution matrix. Recently, MC has been used in different applications in information theory [15–17], information-theoretic security and privacy [18–20], and data processing [21, 22].

Many modern applications include large number of variables with possibly nonlinear relationships among them. Using MC to capture pairwise associations can cause significant over-fitting issues because each variable can be assigned to multiple nonlinear relations. Here we propose *Network Maximal Correlation* (NMC) as a fundamental measure to capture nonlinear associations in networks without the knowledge of underlying nonlinearity shapes. In the NMC optimization, each variable is assigned to at most one transformation function with zero mean and unit variance. NMC infers optimal transformations of variables by maximizing their inner products over edges of the underlying graph. For the case of two variables, NMC is equivalent to MC. The NMC definition does not assume a specific relationship among node variables and the graph structure. For illustration, we consider this relationship in different NMC applications such as graphical model inference. Furthermore, the NMC optimization can be regularized to have even fewer nonlinear transformations to avoid over-fitting issues.

In this paper, we characterize a solution of the NMC optimization using geometric properties of Hilbert spaces for both discrete and continuous jointly Gaussian variables. For discrete random variables, we show that the NMC optimization is an instance of the Maximum Correlation Problem (MCP) which is NP-hard [23–26]. In this case, using results of the Multivariate Eigenvector Problem (MEP) [23], we provide necessary conditions for a global NMC optimum. We also propose an efficient algorithm based on Alternating Conditional Expectation (ACE) [13], which converges to a local NMC optimum. We also provide guidelines for choosing appropriate starting points of the algorithm to jump out of local maximizers. The proposed ACE algorithm does not require forming joint distribution matrices which could be expensive for variables with large alphabet sizes. We also propose a distributed version of the ACE algorithm to compute a 1-ϵ approximation of the NMC value for large and dense graphs using graph partitioning.

For jointly Gaussian variables, we use projections over Hermitte-Chebychev polynomials to characterize an optimal solution of the NMC optimization. Under some conditions, we show that the NMC optimization is equivalent to the Max-Cut problem, which is NP-complete [27]. However, there exist algorithms to approximate its solution using Semidefinte Programming (SDP) within an approximation factor of 0.87856 [28]. In this case, we provide conditions under which an NMC solution can be computed exactly. Using these results, under some general conditions we show that NMC can infer the underlying graphical model for functions of latent jointly Gaussian variables. These functions are unknown, bijective, and can be nonlinear. This result broadens the family of continuous distributions whose graphical models can be characterized efficiently.

In real-world applications, often only noisy samples of joint distributions are available. For this case, we prove a finite sample generalization bound, and error guarantees for NMC. In particular, under general conditions we prove that NMC is continuous with respect to joint probability distributions. That is, a small perturbation in the distribution results in a small change in the NMC value. Moreover, we show that *Sample NMC* (i.e., NMC computed using empirical distributions) converges exponentially fast to the NMC value as the sample size grows.
Moreover, we use the NMC optimization to characterize a nonlinear global relevance graph with a certain complexity [29] and propose a greedy algorithm to infer such a nonlinear relevance graph approximately. Finally, we apply NMC to different cancer datasets [30] including breast, kidney and liver cancers and show that using the NMC network, we can infer gene modules that are significantly associated with survival times of individuals while they are not detected using linear association measures.

2 Maximal Correlation

In this section, we introduce notations and review prior work on maximal correlation.

2.1 Notation

Suppose X_1 and X_2 are two random variables defined on probability space $(\Omega, \mathcal{F}, P)$ taking values in $(\mathcal{X}_1, \mathcal{B}_1)$ and $(\mathcal{X}_2, \mathcal{B}_2)$, respectively. The map $X_i : (\Omega, \mathcal{F}) \to (\mathcal{X}_i, \mathcal{B}_i)$ generates the subalgebra $\mathcal{F}_i = X_i^{-1}(\mathcal{B}_i)$ of $\mathcal{F}$. Let P_{X_i} be the restriction of the measure P on $\mathcal{F}_i$, $i = 1, 2$. For discrete variables, $\mathcal{X}_1$ and $\mathcal{X}_2$ are their finite support sets with cardinalities $|\mathcal{X}_i|$, for $i = 1, 2$.

2.2 Definition and General Properties

A Pearson's linear correlation coefficient between real-valued variables X_1 and X_2 is defined as

$$cor(X_1, X_2) = \frac{\mathbb{E}\left[(X_1 - \mathbb{E}[X_1])\ (X_2 - \mathbb{E}[X_2])\right]}{\sqrt{\text{var}(X_1)}\ \sqrt{\text{var}(X_2)}},$$

where $\text{var}(X_i)$ represents the variance of random variable X_i, for $i = 1, 2$. Correlation-based measures capture linear associations between variables, ignoring possible nonlinear relationships.

Example 1 Suppose X_1 is a Gaussian variable with zero mean and unit variance. Let $X_2 = X_1^2$. In this case, even though variables are strongly associated with each other, the correlation coefficient between them is close to zero (see e.g. Figure 1-a). It is because these variables are related through a nonlinear transformation. One way to capture such a nonlinear relationship between these variables is to quantify maximum correlation between their, possibly nonlinear, transformations. In this example, suppose $\phi_1(X_1) = \alpha_{11} X_1^2 + \alpha_{12}$ and $\phi_2(X_2) = \alpha_{21} X_2 + \alpha_{22}$, where coefficients α_{ij} are selected so that $\phi_i(X_i)$ has zero mean and unit variance, for both $i = 1, 2$. In this case, the correlation coefficient between transformed variables is one (see e.g. Figure 1-b), capturing a strong nonlinear association between variables X_1 and X_2.

Maximal correlation (MC) between variables X_1 and X_2 which was introduced by Gebelein [8] captures a nonlinear association between them by selecting, possibly nonlinear, transformation functions $\phi_1(X_1)$ and $\phi_2(X_2)$ so that $\phi_1(X_1)$ and $\phi_1(X_1)$ have the highest correlation among all other transformation functions with zero means and unit variances.

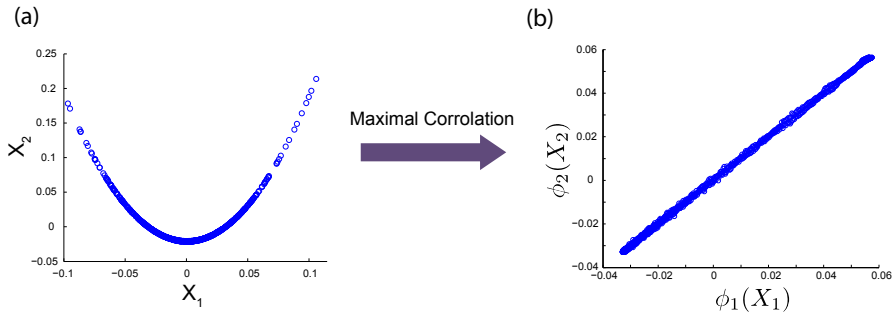

Figure 1: (a) Samples of variables X_1 and X_2 with a nonlinear relationship considered in Example 1. (b) $\phi_1(X_1)$ and $\phi_2(X_2)$ are transformed variables, capturing the nonlinear relationship between X_1 and X_2.

Definition 1 (Maximal Correlation) Maximal correlation between two random variables X_1 and X_2 is defined as

$$\rho(X_1, X_2) \triangleq \max_{\phi_1, \phi_2} \mathbb{E}[\phi_1(X_1)\ \phi_2(X_2)], \tag{2.1}$$

subject to $\phi_i(X_i) : \Omega \to \mathbb{R}$ is measurable[1] , $\mathbb{E}[\phi_i(X_i)] = 0$, and $\mathbb{E}[\phi_i(X_i)^2] = 1$, for $i = 1, 2$.

For $i = 1, 2$, let $\phi_i^*(X_i)$ denote an optimal solution of (2.1). Maximal correlation $\rho(X_1, X_2)$ is always between 0 and 1, where a high MC value indicates a strong association between two variables [8]. The study of maximal correlation and other principle inertia components between two variables dates back to Hirschfeld [9], Gebelein [8], Sarmanov [10], Rényi [11], and Greenacre [12]. Recently, MC has been used in information theory and applied probability problems such as data processing, inference of common randomness among others [10, 14, 22, 31, 32]. Unlike linear correlation, MC only depends on the joint distribution of variables $P_{X_1,X_2}(\cdot,\cdot)$, and not on their alphabets $\mathcal{X}_i$. Several works have investigated different aspects of optimization (2.1) for both discrete [13] and continuous [14] random variables. In particular, the existence of an optimal solution for the MC optimization and the uniqueness of such solutions have been investigated in [13]. Reference [14] has used projections over Hilbert spaces to compute MC for Gaussian variables. We extend this approach to derive existing MC results for discrete variables. In the next section, we use a similar approach based on Hilbert projections to characterize network maximal correlation for both discrete and jointly Gaussian variables.

Definition 2 For $i = 1, 2$, we define a Hilbert space H_i as

$$H_i = \{\phi_i(X_i) | \phi_i(X_i) \text{ is measurable, } \mathbb{E}[\phi_i(X_i)] = 0,\ \mathbb{E}[(\phi_i(X_i))^2] < \infty\},$$

where the product is defined as $\langle \phi_i, \phi_i' \rangle \triangleq \mathbb{E}[\phi_i(X_i)\ \phi_i'(X_i)]$.

Since every Hilbert space has an orthonormal basis (Theorem 2.4, [33]), we let $\{\psi_{1,i}\}_{i=1}^{\infty}$ and $\{\psi_{2,i}\}_{i=1}^{\infty}$ be corresponding orthonormal bases of H_1 and H_2, respectively. Consider the following optimiza-

[1] ϕ_i is a mapping from $\mathcal{X}_i$ to $\mathbb{R}$ and X_i is a mapping from Ω to $\mathcal{X}_i$. Thus, we have $\phi_i(X_i) = \phi_i \circ X_i : \Omega \to \mathbb{R}$.

tion:

$$\begin{aligned} \max_{a_{i,j}} \quad & \sum_{i,j} a_{1,i} a_{2,j}\ \rho_{ij} \\ & \sum_{j=1}^{\infty} a_{i,j}^2 = 1, \quad i = 1,2, \\ & \sum_{j=1}^{\infty} a_{i,j}\ \mathbb{E}[\psi_{i,j}(X_i)] = 0, \quad i = 1,2, \end{aligned} \tag{2.2}$$

where $\rho_{ij} \triangleq \mathbb{E}[\psi_{1,i}(X_1)\ \psi_{2,j}(X_2)]$.

Proposition 1 *Suppose $\phi_i^*(\cdot)$ and $a_{i,j}^*$ are optimal solutions of optimizations (2.1) and (2.2), respectively. Then, we have*

$$\phi_i^*(x) = \sum_{j=1}^{\infty} a_{i,j}^*\ \psi_{i,j}(x). \tag{2.3}$$

Moreover, the joint probability distribution can be written as

$$P_{X_1X_2}(x_1, x_2) = \sum_{i,j} \rho_{ij} \psi_{1,i}(x_1) \psi_{2,j}(x_2).$$

Proof A proof is presented in Section 10.1. ■

Proposition 1 provides an alternative optimization (2.2) to solve the maximal correlation problem (2.1). Selecting appropriate orthonormal bases for Hilbert spaces H_1 and H_2 is critical to obtaining a tractable optimization (2.2). In the following, we use Proposition 1 to solve the maximal correlation optimization for general discrete variables as well as for jointly Gaussian variables.

Example 2 (MC for Discrete Random Variables) Suppose X_1 and X_2 are two discrete random variables with a joint probability function $P_{X_1,X_2}(\cdot,\cdot)$. Let $\{1,\ldots,|\mathcal{X}_1|\}$ and $\{1,\ldots,|\mathcal{X}_2|\}$ be alphabets of random variables X_1 and X_2, respectively. We choose the following orthonormal bases for H_1 and H_2:

$$\psi_{1,i}(x) = \mathbf{1}\{x = i\} \frac{1}{\sqrt{P_{X_1}(i)}} \quad \textit{and} \quad \psi_{2,j}(x) = \mathbf{1}\{x = j\} \frac{1}{\sqrt{P_{X_2}(j)}}.$$

By these selections of bases, we have

$$\rho_{ij} = \mathbb{E}[\psi_{1,i}(X_1)\ \psi_{2,j}(X_2)] = \frac{P_{X_1X_2}(i,j)}{\sqrt{P_{X_1}(i)}\sqrt{P_{X_2}(j)}}.$$

Moreover, we have

$$\mathbb{E}[\psi_{i,j}(X_i)] = \sqrt{P_{X_i}(j)}, \quad i = 1,2.$$

Thus, optimization (2.2) is simplified to the following optimization:

$$\begin{aligned} \max \quad & \sum_{i,j} a_{1,i} a_{2,j} \frac{P_{X_1,X_2}(i,j)}{\sqrt{P_{X_1}(i)}\sqrt{P_{X_2}(j)}} \\ & \sum_{j=1}^{|\mathcal{X}_i|} (a_{i,j})^2 = 1, \quad i = 1,2, \\ & \sum_{j=1}^{|\mathcal{X}_i|} a_{i,j}\ \sqrt{P_{X_i}(j)} = 0, \quad i = 1,2. \end{aligned} \tag{2.4}$$

According to Proposition 1, to solve MC optimization (2.1) it is sufficient to find an optimal solution of optimization (2.4). In the following, we show that an optimal solution of optimization (2.4) can be computed in a closed form using matrix spectral decomposition. Define the normalized joint distribution matrix as

$$Q(i,j) \triangleq \frac{P_{X_1,X_2}(i,j)}{\sqrt{P_{X_1}(i)}\sqrt{P_{X_2}(j)}} \tag{2.5}$$

whose size is $|\mathcal{X}_1| \times |\mathcal{X}_2|$. Let

$$\mathbf{a}_1 \triangleq (a_{1,1}, a_{1,2}, \ldots, a_{1,|\mathcal{X}_1|})^T \quad \textit{and} \quad \mathbf{a}_2 \triangleq (a_{2,1}, a_{2,2}, \ldots, a_{2,|\mathcal{X}_2|})^T$$

be coefficient vectors. Moreover, let

$$\begin{aligned} \sqrt{\mathbf{p}_1} &\triangleq \left(\sqrt{P_{X_1}(1)}, \sqrt{P_{X_1}(2)}, \ldots, \sqrt{P_{X_1}(|\mathcal{X}_1|)}\right)^T \\ \sqrt{\mathbf{p}_2} &\triangleq \left(\sqrt{P_{X_2}(1)}, \sqrt{P_{X_2}(2)}, \ldots, \sqrt{P_{X_2}(|\mathcal{X}_2|)}\right)^T \end{aligned} \tag{2.6}$$

be vectors of square roots of marginal probabilities. Optimization (2.4) can be re-written as follows:

$$\begin{aligned} \max \quad & \mathbf{a}_1^T \, Q \, \mathbf{a}_2 \\ & \|\mathbf{a}_i\|_2 = 1, \quad i = 1,2, \\ & \mathbf{a}_i \perp \sqrt{\mathbf{p}_i}, \quad i = 1,2. \end{aligned} \tag{2.7}$$

In the following, we show that optimal coefficient vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ of optimization (2.7) are equal to the left and right singular vectors of the matrix Q corresponding to its second largest singular value. Moreover, the optimal value (the maximal correlation between two variables X_1 and X_2) is equal to the second largest singular value of the matrix Q. To show this, we define random variables Z_1 and Z_2 such that

$$\mathbb{P}\left[Z_1 = \frac{a_{1,i}}{\sqrt{P_{X_1}(i)}}, Z_2 = \frac{a_{2,j}}{\sqrt{P_{X_2}(j)}}\right] = P_{X_1,X_2}(i,j),$$

where $\|\mathbf{a}_1\| = 1$ and $\|\mathbf{a}_2\| = 1$. Using the Cauchy-Schwartz inequality, we have that

$$\mathbf{a}_1^T Q \mathbf{a}_2 = \mathbb{E}[Z_1 Z_2] \leq \sqrt{\mathbb{E}[Z_1^2]\mathbb{E}[Z_2^2]} = \|\mathbf{a}_1\| \, \|\mathbf{a}_2\| = 1.$$

Therefore, the maximum singular value of Q is at most one. Using (2.5), one can see that the right and left singular vectors of Q with the singular value one are $\sqrt{\mathbf{p}_1}$ and $\sqrt{\mathbf{p}_2}$, respectively. Thus, the feasible set of optimization (2.7) includes unit-norm vectors orthogonal to leading singular vectors of Q. Thus, the optimal value is equal to the second largest singular value and optimal vectors $\mathbf{a}_1^*$ and $\mathbf{a}_2^*$ are left and right singular vectors corresponding to the second largest singular value.

Example 3 (MC for Jointly Gaussian Random Variables) This example is studied in reference [14] to compute MC between two Gaussian variables. In Section 6, we use a similar approach to characterize network maximal correlation for jointly Gaussian variables.
Suppose (X_1, X_2) are jointly Gaussian variables with the correlation coefficient ρ. The k-th Hermitte-chebychev polynomial is defined as

$$\Psi_k(x) = (-1)^k e^{x^2} \frac{d^k}{dx^k} e^{-x^2}. \tag{2.8}$$

These polynomials form an orthonormal basis with respect to Gaussian distributions. That is,

$$\int_{-\infty}^{\infty} H_i(x_1)H_j(x_2)f(x_1,x_2)dx_1dx_2 = \rho^i \mathbb{1}_{i=j}, \tag{2.9}$$

where $f(x,y)$ is the joint density function of Gaussian variables with correlation ρ, and $\mathbb{1}_{i=j}$ is one when $i=j$, otherwise it is zero. Let $\psi_{i,j}$ to be the j-th Hermitte-Chebychev polynomial, for $i=1,2$. Using (2.9), we have

$$\rho_{ij} = \mathbb{E}[\psi_{1,i}(X_1)\ \psi_{2,j}(X_2)] = \rho^i \mathbb{1}_{i=j}.$$

Moreover, we have

$$\mathbb{E}[\psi_{i,j}(X_i)] = \mathbb{1}_{j=0}, \quad i=1,2, \tag{2.10}$$

because all of these functions for $j \geq 1$ have zero means over a Gaussian distribution. Therefore, optimization (2.2) can be written as

$$\begin{aligned} \max \quad & \sum_{i=0}^{\infty} a_{1,i}a_{2,i}\rho^i \\ & \sum_{j=0}^{\infty} (a_{i,j})^2 = 1, \quad i=1,2, \\ & a_{i,0} = 0, \quad i=1,2. \end{aligned} \tag{2.11}$$

Since $|\rho| \leq 1$, an optimal solution of optimization (2.11) is obtained when $|a_{1,1}|=1$, $|a_{2,1}|=1$, while other coefficients are equal to zero. The signs of $a_{i,1}$ for $i=1,2$ are determined so that $a_{1,1}a_{2,1}\rho = |\rho|$. This leads to the maximal correlation $|\rho|$ between two variables that is equal to the absolute value of the correlation coefficient between them when the two random variables are jointly Gaussian. Moreover, optimal transformation functions are

$$\phi_i(X_i) = a_{i,1}\psi_{i,1} = \pm X_i, \quad i=1,2,$$

where signs of variables are selected so that $a_{1,1}a_{2,1}\rho = |\rho|$.

3 Statistical Properties of Maximal Correlation

In many applications, often only noisy samples of joint distributions are observed. In this section, we prove a finite sample generalization bound, and error guarantees for maximal correlation of discrete random variables. Specifically, under some general conditions, we prove that

- maximal correlation is a continuous measure with respect to the joint probability distribution. That is, a small perturbation in the distribution results in a small change in the MC value.
- sample maximal correlation between two variables, computed using m samples from the joint distribution, converges exponentially fast to the MC value, as m grows.

These properties establish maximal correlation as a robust association measure to capture nonlinear dependencies between variables in real-world applications.

Throughout this subsection we only consider discrete random variables and assume that all alphabet elements $x_i \in \mathcal{X}_i$ have positive probabilities (otherwise they can be neglected without loss of generality). That is, if

$$\delta_i \triangleq \arg\min_{x_i \in \mathcal{X}_i} P_{X_i}(x_i), \quad i = 1, 2, \tag{3.1}$$

then $\delta(P) \triangleq \min\{\delta_1, \delta_2\} > 0$. The empirical distribution of these variables using m observed samples is defined as $P^{(m)}(x_1, x_2) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{x_1^{(i)} = x_1, x_2^{(i)} = x_2\}$, where $\{x_1^{(i)}, x_2^{(i)}\}_{i=1}^{m}$ are i.i.d. samples drawn according to a distribution P_{X_1,X_2}. The vector of observed samples of variable X_i is denoted by $\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)})$. For any vector $\mathbf{v} = (v_1, \ldots, v_p) \in \mathbb{R}^d$ and $p \geq 1$, we let $\|\mathbf{v}\|_p$ represent the standard p-norm of the vector $\mathbf{v}$ defined as

$$\|\mathbf{v}\|_p = \left(\sum_{i=1}^{d} v_i^p \right)^{\frac{1}{p}}.$$

For $p = 2$, we drop the subscript if no confusion arises, i.e., $\|\mathbf{v}\| = \|\mathbf{v}\|_2$.

3.1 Continuity of Maximal Correlation

Let $P_{X_1,X_2}(\cdot,\cdot)$ and $\tilde{P}_{X_1,X_2}(\cdot,\cdot)$ be two distributions over alphabets $(\mathcal{X}_1, \mathcal{X}_2)$ with the corresponding MC values ρ and $\tilde{\rho}$, respectively. In the following, we show that if the distance between P and $\tilde{P}$ is small (i.e., $\|P - \tilde{P}\|_\infty \leq \epsilon$), their corresponding MC values (ρ and $\tilde{\rho}$) are close to each other as well.

Theorem 1 *Let $\|P - \tilde{P}\|_\infty \leq \epsilon$, for some $\epsilon > 0$. Then, we have*

$$|\rho - \tilde{\rho}| \leq 2 \frac{\epsilon}{\delta^2} D^{3/2}, \tag{3.2}$$

where $D \triangleq \max\{|\mathcal{X}_1|, |\mathcal{X}_2|\}$, and $\delta \triangleq \min(\delta(P), \delta(\tilde{P}))$.

Proof A proof is presented in Section 10.2. ∎

The sketch of the proof is as follows: The normalized joint distribution matrix Q (2.5) can be written as

$$Q = D_{X_1}(P)^{-\frac{1}{2}} P_{X_1,X_2} D_{X_2}(P)^{-\frac{1}{2}}, \tag{3.3}$$

where $D_{X_i}(P)$ denotes a diagonal matrix whose diagonal is P_{X_i}, for $i = 1, 2$. Since the matrix Q is a continuous function of $P_{X_1,X_2}(\cdot,\cdot)$, its singular values (and therefore its second largest singular value) are continuous functions of P as well.

3.2 Sample Maximal Correlation

Let $\{x_1^i, x_2^i\}_{i=1}^{m}$ be i.i.d. samples drawn according to a joint probability distribution $P_{X_1,X_2}(\cdot,\cdot)$. Suppose $P^{(m)}(\cdot,\cdot)$ denotes the empirical distribution obtained from these samples. Maximal correlation computed using this empirical probability distribution is called *Sample Maximal Correlation* and is denoted by $\rho_m(X_1, X_2)$. In the following, we show that $\rho_m(X_1, X_2)$ converges to $\rho(X_1, X_2)$ exponentially fast, as $m \to \infty$.

Theorem 2 *For any distribution P, and any $\epsilon > 0$, $\mathbb{P}\left[|\rho_m(X_1, X_2) - \rho(X_1, X_2)| > \epsilon\right] \to 0$, exponentially fast. More precisely, if*

$$m \geq \frac{3}{\delta(P)^2 \epsilon} \sqrt{D} \log\left(\frac{24}{\eta}\right),$$

then,

$$\mathbb{P}\left[|\rho_m(X_1, X_2) - \rho(X_1, X_2)| > \epsilon\right] \leq \eta,$$

where $D = \max\{|\mathcal{X}_1|, |\mathcal{X}_2|\}$. The bound can also be written as

$$\mathbb{P}\left[|\rho_m(X_1, X_2) - \rho(X_1, X_2)| > \epsilon\right] \leq \frac{1}{24} \exp\left(-m \frac{\delta(P)^2 \epsilon}{3\sqrt{D}}\right).$$

Proof A proof is presented in Section 10.3. ■

The proof follows from the facts that maximal correlation is a continuous function of the input distribution according to Theorem 1, and the empirical distribution converges exponentially fast to the true distribution.

4 Network Maximal Correlation

4.1 Definition and General Properties

In this section, we introduce *Network Maximal Correlation* (NMC) as a fundamental measure to capture nonlinear associations over networks. Let $G = (V, E)$ be a graph with n nodes and $|E|$ edges. The graph G is un-weighted, does not have self-loops, and can be directed or undirected. Each node i is assigned to a random variable X_i. Here, we introduce NMC without assuming a specific relationship among node variables and the graph structure. We discuss this relationship in different applications of NMC in Sections 6, 7, and 8. NMC infers best nonlinear transformation functions assigned to each node variable so that the total pairwise correlation over the network is maximized.

Suppose $X_1, \ldots, X_n$ are n random variables defined on probability space $(\Omega, \mathcal{F}, P)$, where X_i takes values in $(\mathcal{X}_i, \mathcal{B}_i)$. The map $X_i : (\Omega, \mathcal{F}) \to (\mathcal{X}_i, \mathcal{B}_i)$ generates the subalgebra $\mathcal{F}_i = X_i^{-1}(\mathcal{B}_i)$ of $\mathcal{F}$. Let P_{X_i} be the restriction of the measure P on $\mathcal{F}_i$, $i = 1, \ldots, n$. For discrete variables, $\mathcal{X}_1$ and $\mathcal{X}_2$ are their finite support sets with cardinalities $n_i = |\mathcal{X}_i|$, for $i = 1, \ldots, n$.

Definition 3 (Network Maximal Correlation) Network maximal correlation among variables $X_1, \ldots, X_n$ connected by a graph $G = (V, E)$ is defined as

$$\rho_G(X_1, \ldots, X_n) \triangleq \max_{\phi_1, \ldots, \phi_n} \sum_{(i,j) \in E} \mathbb{E}[\phi_i(X_i)\ \phi_j(X_j)], \tag{4.1}$$

subject to $\phi_i(X_i) : \Omega \to \mathbb{R}$ is measurable, $\mathbb{E}[\phi_i(X_i)] = 0$, and $\mathbb{E}[\phi_i(X_i)^2] = 1$, for $1 \leq i \leq n$.

The Optimization (4.1) maximizes total pairwise correlation over the network without distinguishing among positive and negative correlations. In some applications, the strength of an association does not depend on the sign of the correlation coefficient. In those cases, one can re-write the NMC optimization (4.1) to maximize the total absolute pairwise correlations over the network as follows:

Definition 4 (Absolute Network Maximal Correlation) Consider the following optimization:

$$\max_{\phi_1,\dots,\phi_n} \sum_{(i,j)\in E} |\mathbb{E}[\phi_i(X_i)\ \phi_j(X_j)]|, \tag{4.2}$$

subject to $\phi_i(X_i) : \Omega \to \mathbb{R}$ is measurable, $\mathbb{E}[\phi_i(X_i)] = 0$ and $\mathbb{E}[\phi_i^2(X_i)] = 1$, for any $1 \le i \le n$. We refer to this optimization as an absolute NMC optimization.

Let $\phi_i^*(\cdot)$ be an optimal solution of the NMC optimization (4.1) (in Proposition 2 we prove the existence of such solution). Then, an edge maximal correlation between variables i and j is defined as

$$\rho_G(X_i, X_j) \triangleq \big|\mathbb{E}[\phi_i^*(X_i)\ \phi_j^*(X_j)]\big|, \tag{4.3}$$

where $(i,j) \in E$. Unlike maximal correlation formulation of (2.1), transformation functions in optimization (4.1) are constrained by the network structure. Therefore, an edge maximal correlation between variables X_i and X_j is always smaller than or equal to their maximal correlation, i.e.,

$$\rho_G(X_i, X_j) \le \rho(X_i, X_j).$$

Computation of maximal correlation for each edge independently results in two nonlinear functions assigned to nodes of that edge. Therefore, if the network has $|E|$ edges, it will result in inference of $2|E|$ possibly nonlinear functions. In that setup, each node can be associated to different nonlinear transformation functions which can raise over-fitting issues particularly for dense networks. On the other hand, in the NMC formulation of (4.1), we assign a *single* function to each node in the graph. Therefore, optimization (4.1) results in n possibly nonlinear functions.

Lemma 1 The NMC optimization (4.1) is equivalent to the following MSE optimization:

$$\min_{\phi_1,\dots,\phi_n} \ \frac{1}{2} \sum_{(i,j)\in E} \mathbb{E}[(\phi_i(X_i) - \phi_j(X_j))^2], \tag{4.4}$$

where $\mathbb{E}[\phi_i(X_i)] = 0$ and $\mathbb{E}[\phi_i^2(X_i)] = 1$, for any $1 \le i \le n$.

Proof A proof is presented in Section 10.4. ■

Similarly to Definition 2, for $i = 1, 2, \dots, n$, we define a Hilbert space H_i as

$$H_i = \{\phi_i(X_i) | \phi_i(X_i) \text{ is measurable, } \mathbb{E}[\phi_i(X_i)] = 0,\ \mathbb{E}[(\phi_i(X_i))^2] < \infty\},$$

where the product is defined as $\langle \phi_i, \phi_i' \rangle \triangleq \mathbb{E}[\phi_i(X_i)\ \phi_i'(X_i)]$.
The following proposition shows the existence of optimal transformations of the NMC optimization (4.1):

Proposition 2 *Under the assumption that Hilbert spaces H_i's are compact, there exist functions ϕ_i^* such that $\mathbb{E}[\phi_i^*(X_i)] = 0$ and $\mathbb{E}[\phi_i^*(X_i)^2] = 1$ for $1 \le i \le n$, that achieve the optimal value of optimization (4.1).*

Proof A proof is presented in Section 10.5. ■

The assumption that Hilbert spaces H_i's are compact holds when X_i's are discrete random variables with finite support, or when X_i's are jointly Gaussian random variables.

Let $\mathcal{P}_i$ denote the projection operation from the space H_j (for any $j \neq i$) onto H_i, for any $1 \leq i \leq n$. According to Lemma 5 this projection can be characterized using conditional expectations a s follows: For random variable $\phi_j \in H_j$, we have

$$\mathcal{P}_i \phi_j = \text{argmin}_{\phi_i \in H_i} \mathbb{E}\left[(\phi_i - \phi_j)^2\right] = \frac{\mathbb{E}[\phi_j | X_i]}{\sqrt{\mathbb{E}[\phi_j | X_i]^2}}.$$

The following proposition characterizes optimal NMC transformation functions using projection operators:

Proposition 3 *Optimal transformation functions of NMC optimization* (4.1) $\{\phi_i^*, 1 \leq i \leq n\}$ *satisfy*

$$\phi_i^* = \frac{\sum_{j \in \mathcal{N}(i)} \mathcal{P}_i \ \phi_j^*}{\| \sum_{j \in \mathcal{N}(i)} \mathcal{P}_i \ \phi_j^* \|} = \frac{\mathbb{E}[\sum_{j \in \mathcal{N}(i)} \phi_j^* | X_i]}{\| \mathbb{E}[\sum_{j \in \mathcal{N}(i)} \phi_j^* | X_i] \|}, \tag{4.5}$$

where $\mathcal{N}(i)$ *represents neighbors of node* i *in the graph* $G = (V, E)$.

Proof A proof is presented in Section 10.6. ■

Note 1 A similar approach can be used to characterize the absolute NMC optimization (4.2) by introducing extra variables to represent correlation signs of edges:

$$\begin{aligned} \max \quad & \sum_{(i,j) \in E} s_{i,j} \mathbb{E}[\phi_i(X_i) \phi_j(X_j)] \\ & \mathbb{E}[\phi_i(X_i)] = 0, \ \mathbb{E}[\phi_i^2(X_i)] = 1, \quad 1 \leq i \leq n. \end{aligned} \tag{4.6}$$

In this case, similarly to Proposition (10.6), we can write

$$\phi_i^* = \frac{\sum_{j \in \mathcal{N}(i)} s_{ij}^* \mathcal{P}_i \ \phi_j^*}{\| \sum_{j \in \mathcal{N}(i)} s_{ij}^* \mathcal{P}_i \ \phi_j^* \|},$$

where

$$s_{ij}^* = \text{sign}\left(\mathbb{E}[\phi_i^*(X_i) \phi_j^*(X_j)]\right).$$

Proposition 3 characterizes a property of optimal transformations of NMC (4.1) using projection operations without explicitly computing the optimal NMC solution. In the following, we use orthonormal representations of the Hilbert spaces H_i and propose a constructive approach to solve the NMC optimization.
Recall that $\{\psi_{i,j}\}_{j=1}^{\infty}$ represents an orthonormal basis for H_i. Consider the following optimization:

$$\begin{aligned} \max \quad & \sum_{(i,i') \in E} \sum_{j,j'} a_{i,j} a_{i',j'} \ \rho_{i,i'}^{j,j'} \\ & \sum_{j=1}^{\infty} a_{i,j}^2 = 1, \quad 1 \leq i \leq n, \\ & \sum_{j=1}^{\infty} a_{i,j} \ \mathbb{E}[\psi_{i,j}(X_i)] = 0, \quad 1 \leq i \leq n, \end{aligned} \tag{4.7}$$

where $\rho_{i,i'}^{j,j'} \triangleq \mathbb{E}[\psi_{i,j}(X_i) \ \psi_{i',j'}(X_{i'})]$.

Theorem 3 *Suppose $\phi_i^*(\cdot)$ and $a_{i,j}^*$ are optimal solutions of optimizations (4.1) and (4.7), respectively. Then, we have*

$$\phi_i^*(x) = \sum_{j=1}^{\infty} a_{i,j}^* \ \psi_{i,j}(x). \tag{4.8}$$

Proof A proof is presented in Section 10.7. ■

Similarly to the case of two variables discussed in Section 2.2, selecting appropriate Hilbert spaces H_i is critical to have a tractable optimization (4.7). In the following, we consider the NMC optimization for discrete variables, while the Gaussian case is discussed in Section 6.

Example 4 (NMC for Discrete Random Variables) Suppose X_i is a discrete random variable with alphabet $\{1, \ldots, |\mathcal{X}_i|\}$. Similarly to Example 2, let $\psi_{i,j}(x) = \mathbf{1}\{x = j\}\frac{1}{\sqrt{P_{X_i}(j)}}$ be an orthonormal basis for H_i. Thus, we have

$$\rho_{i,i'}^{j,j'} = \mathbb{E}[\psi_{i,j}(X_i) \ \psi_{i',j'}(X_{i'})] = \frac{P_{X_iX_{i'}}(j,j')}{\sqrt{P_{X_i}(j)}\sqrt{P_{X_{i'}}(j')}}.$$

Therefore, optimization (4.7) is simplified to the following optimization:

$$\begin{aligned} \max \quad & \sum_{(i,i')\in E} \sum_{j,j'} a_{i,j}a_{i',j'} \ \frac{P_{X_iX_{i'}}(j,j')}{\sqrt{P_{X_i}(j)}\sqrt{P_{X_{i'}}(j')}} \\ & \sum_{j=1}^{|\mathcal{X}_i|} (a_{i,j})^2 = 1, \quad 1 \le i \le n, \\ & \sum_{j=1}^{|\mathcal{X}_i|} a_{i,j} \ \sqrt{P_{X_i}(j)} = 0, \quad 1 \le i \le n. \end{aligned} \tag{4.9}$$

Similarly to Example 2, we define the matrix $Q_{i,i'}$ as

$$Q_{i,i'}(j,j') \triangleq \frac{P_{X_i,X_{i'}}(j,j')}{\sqrt{P_{X_i}(j)}\sqrt{P_{X_{i'}}(j')}}, \tag{4.10}$$

whose size is $|\mathcal{X}_i| \times |\mathcal{X}_{i'}|$. Moreover, recall that for $i = 1, \ldots, n$, we have

$$\begin{aligned} \mathbf{a}_i &= (a_{i,1}, a_{i,2}, \ldots, a_{i,|\mathcal{X}_i|})^T \\ \sqrt{\mathbf{p}_i} &= \left(\sqrt{P_{X_i}(1)}, \sqrt{P_{X_i}(2)}, \ldots, \sqrt{P_{X_i}(|\mathcal{X}_i|)}\right)^T. \end{aligned}$$

Therefore, optimization (4.9) can be re-written as follows:

$$\begin{aligned} \max \quad & \sum_{(i,i')\in E} \mathbf{a}_i^T Q_{i,i'} \mathbf{a}_{i'} \\ & \|\mathbf{a}_i\|_2 = 1, \quad 1 \le i \le n, \\ & \mathbf{a}_i \perp \sqrt{\mathbf{p}_i}, \quad 1 \le i \le n. \end{aligned} \tag{4.11}$$

Optimization (4.11) is not convex nor concave in general. In Section 5.2, we show that this optimization is an instance of the standard Maximum Correlation Problem (MCP) proposed by Hotelling [24, 25]. By making this connection, we use established techniques of solving Multivariate Eigenvalue Problem (MEP) to solve optimization (4.11).

4.2 Statistical Properties of NMC

In this part, we investigate the robustness of NMC for discrete variables with finite support against small perturbations of joint probability distributions of variable pairs. Moreover, we show that sample NMC (i.e., NMC computed using empirical distributions) converges to the true NMC value exponentially fast as the sample size increases. To simplify notation, suppose $P_{i,i'}$ is the matrix representation of the joint probability distribution of variables X_i and $X_{i'}$.

Theorem 4 Network maximal correlation is a continuous function of the joint probability distributions $P_{i,i'}$, for all $(i,i') \in E$. Let $\|P_{i,i'} - \tilde{P}_{i,i'}\|_\infty \leq \epsilon$, for some $\epsilon > 0$, and all $(i,i') \in E$. Then, we have

$$|\tilde{\rho}_G - \rho_G| \leq \epsilon |E| D^{\frac{3}{2}} \frac{6}{\delta^2}, \tag{4.12}$$

where $D = \max\{|\mathcal{X}_1|, \ldots, |\mathcal{X}_n|\}$, and $\delta = \min_{1\leq i\leq n} \big(\min\{\delta(P_{X_i}), \delta(\tilde{P}_{X_i})\}\big)$.

Proof A proof is presented in Section 10.8. ∎

Next, we show that the sample NMC denoted by $\rho_m(G)$ converges to the true NMC value ρ_G exponentially fast, as the sample size m increases:

Theorem 5 *Sample NMC converges to NMC, exponentially fast. Particularly, let* $\delta = \min_{1\leq i\leq n} \delta(P_{X_i})$ *and* $D = \max\{|\mathcal{X}_1|, \ldots, |\mathcal{X}_n|\}$. *Then, for*

$$m \geq \left(\frac{24|E|^2 D^3}{\epsilon^2 \delta^2}\right) \log\left(\frac{8\max\{|V|,|E|\}}{\eta}\right), \tag{4.13}$$

we have

$$\mathbb{P}[|\rho_m(G) - \rho_G| > \epsilon] \leq \eta. \tag{4.14}$$

Proof A proof is presented in Section 10.9. ∎

Note that for the case of having two variables, robustness bounds provided in Theorems 4 and 5 are more loose compared to bounds provided by Theorem 1 and 2 owing to the generality of relaxations used in NMC performance characterization.

4.3 Regularized NMC

In this section, we assume that all variables are real-valued (note that this is not a necessary condition for MC and NMC). The NMC optimization (4.1) results in n possibly nonlinear transformation functions $\phi_i^*(X_i)$ whose distances from original variables can be arbitrarily large (i.e., $\mathbb{E}[\phi_i^*(X_i)\ X_i]$ can be arbitrarily small). In some applications, one may wish to have fewer than n nonlinear transformation functions assigned to variables, or alternatively to control distances among transformed and original variables. Here, we propose a regularized NMC optimization framework which penalizes distances among optimal transformation functions $\phi_i^*(X_i)$ and the original variables X_i. Suppose variables have mean zero and unit variance. I.e., $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^2] = 1$.

Algorithm 1 Alternating Conditional Expectation

Initialization: $\phi_1^{(0)}(X_1), \phi_2^{(0)}(X_2)$ with mean zero.
for k=0,1, ...
$\phi_1^{(k+1)}(X_1) = \mathbb{E}[\phi_2^{(k)}(X_2)|X_1]$.
update: $\phi_1^{(k+1)}(X_1) = \frac{\phi_1^{(k+1)}(X_1)}{\sqrt{\mathbb{E}[(\phi_1^{(k+1)}(X_1))^2]}}$
$\phi_2^{(k+1)}(X_2) = \mathbb{E}[\phi_1^{(k)}(X_1)|X_2]$.
update: $\phi_2^{(k+1)}(X_2) = \frac{\phi_2^{(k+1)}(X_2)}{\sqrt{\mathbb{E}[(\phi_2^{(k+1)}(X_2))^2]}}$
update: $\rho^{(k+1)} = \mathbb{E}[\phi_1^{(k+1)}(X_1)\phi_2^{(k+1)}(X_2)]$
end

Definition 5 (Regularized NMC) Regularized NMC among variables $X_1, \ldots, X_n$ connected by a graph $G = (V, E)$ is defined as the solution of the following optimization:

$$\max_{\phi_1,\ldots,\phi_n} (1-\lambda) \sum_{(i,j)\in E} \mathbb{E}[\phi_i(X_i)\ \phi_j(X_j)] + \lambda \sum_{i\in V} \mathbb{E}[\phi_i(X_i)\ X_i], \tag{4.15}$$

where $\mathbb{E}[\phi_i(X_i)] = 0$ and $\mathbb{E}[\phi_i^2(X_i)] = 1$, for any $1 \le i \le n$. $0 \le \lambda \le 1$ is the regularization parameter.

Unlike MC and NMC, which only depend on the joint distributions of variables, the regularized NMC depends on both joint distributions and alphabets of variables because of the regularization term. Moreover, one can define regularized absolute NMC similarly to optimization (4.2).
Let optimal transformation functions computed by optimization (4.15) be $\phi_{i,\lambda}^*$. If $\lambda = 0$, $\phi_{i,\lambda}^* = \phi_i^*$, while if $\lambda = 1$, $\phi_{i,\lambda}^* = X_i$. By varying λ between 0 and 1, transformation functions vary from ϕ_i^* to X_i. Suppose

$$\rho_{G,\lambda}(X_1, \ldots, X_n) \triangleq \sum_{(i,j)\in E} \mathbb{E}[\phi_{i,\lambda}^*(X_i)\ \phi_{j,\lambda}^*(X_j)].$$

Therefore, $\rho_{G,0} = \rho_G$ and $\rho_{G,1}$ is the total linear correlations over the network. By the definition of NMC, $\rho_{G,0} \ge \rho_{G,1}$.

5 Computation of MC and NMC

In this section, we first review an existing algorithm to compute MC and then introduce an efficient algorithm to compute NMC. We also propose a parallelizable version of the NMC algorithm based on network partitioning and show that its expected performance is ϵ-away from the true NMC value.

5.1 Computation of Maximal Correlation

Given the joint distribution of variables, one can use Proposition 1 to compute the MC value and optimal transformation functions. In particular, for discrete random variables, Example 2 shows that a solution of optimization (2.1) can be characterized by the second largest singular value of the normalized joint distribution matrix Q (2.5). Given samples of variables (i.e., $\{x_1^{(i)}, x_2^{(i)}\}_{i=1}^m$), one can compute MC using the empirical joint distribution of variables $P_m(\cdot,\cdot)$. Robustness of MC

computation using empirical distributions is discussed in Section 3. If alphabet sizes of variables are large, forming the joint distribution matrix can be costly. An iterative approach to compute maximal correlation without forming the joint distribution function is based on *Alternating Conditional Expectation (ACE)* [13]. Briefly, at each interaction, the ACE algorithm computes optimal transformation functions using conditional expectations, assuming that the other transformation function is fixed (in a Gauss-Seidel manner [34]). If the correlation value does not increase by a certain value, the algorithm terminates. We describe the steps of this algorithm in Algorithm 1.

Proposition 4 *The sequence $\rho^{(k)}$ generated by Algorithm 1 converges to a local optimum of optimization (2.1). Moreover, if starting points of Algorithm 1 are such that vectors*

$$\left(\phi_1^{(0)}(1)\sqrt{p_{X_1}(1)},\ldots,\phi_1^{(0)}(|\mathcal{X}_1|)\sqrt{p_{X_1}(|\mathcal{X}_1|)}\right)$$

and

$$\left(\phi_2^{(0)}(1)\sqrt{p_{X_2}(1)},\ldots,\phi_2^{(0)}(|\mathcal{X}_2|)\sqrt{p_{X_2}(|\mathcal{X}_2|)}\right)$$

are not orthogonal to the span of the left and right singular vectors corresponding to the second largest singular value of Q, then ACE algorithm 1 converges to the global optimum. Moreover, if the Q matrix has unique singular vectors (left and right) corresponding to the second largest singular value, optimal transformation functions are unique maximizers of optimization (2.1).

Proof See Theorems 5.4 and 5.5 of reference [13]. ■

5.2 Computation of NMC

In this section, we first establish a connection between the NMC optimization (4.1) with Maximum Correlation Problem (MCP) and Multivariate Eigenvalue Problem (MEP) ([23–26]). Then, we deploy techniques used to solve MEP and MCP in order to compute NMC. The Maximum Correlation Problem (MCP), proposed by Hotelling [24, 25], is to find the linear combination of one set of variables that correlates maximally with the linear combination of another set of variables. This problem is defined as

$$\begin{aligned}\max_{\mathbf{b}_i} \quad & \sum_{i,j=1}^{n} \mathbf{b}_i^T C_{i,j}\mathbf{b}_j \\ & \|\mathbf{b}_i\| = 1, \quad 1 \le i \le n,\end{aligned} \tag{5.1}$$

where $\mathbf{b}_i \in \mathbb{R}^{n_i}$ and $C_{i,j} \in \mathbb{R}^{n_i \times n_j}$. Optimization (5.1) is in the standard form of the MCP problem [24,25]. Upon employing the Lagrange multiplier theory [34], the first-order optimality condition for optimization (5.7) is the existence of real-valued scalars, namely, Lagrange multipliers $\lambda_1,\ldots,\lambda_n$, such that the following system of equations is satisfied:

$$\begin{aligned}&\sum_{j=1}^{n} C_{ij}\mathbf{b}_j = \lambda_i \mathbf{b}_i, \quad 1 \le i \le n \\ &\|\mathbf{b}_i\| = 1, \quad 1 \le i \le n.\end{aligned} \tag{5.2}$$

This system of equations is called Multivariate Eigenvalue Problem (MEP). We next establish the connection between NMC and MCP. To that end, we define the following notation: For each i,

since $I_{|\mathcal{X}_i|} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_i}^T$ is positive semidefinite, we take its square root[2] and write

$$I - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_i}^T = B_i B_i^T,$$

where $I_{|\mathcal{X}_i|}$ is a $|\mathcal{X}_i| \times |\mathcal{X}_i|$ identity matrix. Let $\mathbf{b}_i \triangleq B_i \mathbf{a}_i$. Let $U_i \Sigma_i U_i^T$ be the singular value decomposition of B_i where $U_i^{(j)}$ is the j-th column of U_i and $\sigma_i^{(j)}$ is the j-th singular value of B_i. We will show that only one singular value of B_i is zero which is equal to the singular vector $\sqrt{\mathbf{p}_i}$. Without loss of generality, suppose $\sigma_i^1 = 0$, for all i. Define A_i a $|\mathcal{X}_i| \times |\mathcal{X}_i|$ matrix as follows:

$$A_i \triangleq \left(\left[U_i^{(2)}, \ldots, U_i^{(|\mathcal{X}_i)|} \right] \operatorname{diag}\left(\frac{1}{\sigma_i^{(2)}}, \ldots, \frac{1}{\sigma_i^{(|\mathcal{X}_i|)}} \right) \left[U_i^{(2)}, \ldots, U_i^{(|\mathcal{X}_i|)} \right]^T \right). \tag{5.3}$$

Since $\sigma_i^{(j)} \neq 0$, for all $1 \leq i \leq n$, and $j \geq 2$, thus A_i is well-defined according to (5.3).

Theorem 6 *The NMC optimization (4.11) can be re-written as follows:*

$$\max_{\mathbf{b}_i} \quad \sum_{(i,i') \in E} \mathbf{b}_i^T A_i^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_{i'}}^T \right) A_{i'} \mathbf{b}_{i'} \tag{5.4}$$

$$s.t. \ \|\mathbf{b}_i\|_2 = 1. \tag{5.5}$$

Proof A proof is presented in Section 10.10. ∎

Let C be a matrix consisting of submatrices $C_{i,i'}$ where if $(i, i') \in E$,

$$C_{i,i'} \triangleq A_i^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_{i'}}^T \right) A_{i'}, \tag{5.6}$$

otherwise $C_{i,i'}$ is an all zero matrix of size $|\mathcal{X}_i| \times |\mathcal{X}_{i'}|$. Let $\mathbf{b} \triangleq (\mathbf{b}_1^T, \ldots, \mathbf{b}_n^T)^T \in \mathbb{R}^M$, where $\mathbf{b}_i \in \mathbb{R}^{|\mathcal{X}_i|}$ and $M = \sum_{i=1}^n |\mathcal{X}_i|$.

Proposition 5 *The NMC optimization (5.4) can be written as follows:*

$$\begin{aligned} \max \quad & \mathbf{b}^T C \mathbf{b} \\ & \|\mathbf{b}_i\|_2 = 1, \quad 1 \leq i \leq n. \end{aligned} \tag{5.7}$$

Optimization (5.7) is in the standard form of the MCP problem [24, 25]. After showing that the NMC optimization can be reformulated as the MCP, we use the existing techniques in the literature to solve it. Several local maximizers exist for cases that finding a global optimum of optimization (5.7) is computational difficult [23, 35]. For example, an aggregated power method that iterates on blocks of C was proposed by Horst [36] as a general technique for solving the MEP numerically. Below, we summarize general algorithmic ideas to solve MCP:

(1) First, an efficient algorithm is used to solve MEP which is the necessary first order condition for MCP. This step is studied in references [23, 36].

(2) Next, a strategy is used to properly choose starting points of the algorithm or jump out of the local minima of optimization (5.7). This step is studied in [26, 37].

Algorithm 2 Gauss-Seidel Algorithm for MEP

Input: $C \in \mathbb{R}^M \times \mathbb{R}^M$.
Initialization: $\mathbf{b}^{(0)} \in \mathbb{R}^M$.
for $k = 0, 1, \ldots$
  for $i = 1, \ldots, n$
    $\tilde{\mathbf{b}}_i^{(k)} = \sum_{j=1}^{i-1} C_{ij}\mathbf{b}_j^{(k+1)} + \sum_{j=i}^{n} C_{ij}\mathbf{b}_j^{(k)}$.
    $\lambda_i^{(k)} = \|\tilde{\mathbf{b}}_i^{(k)}\|_2$.
    $\mathbf{b}_i^{(k+1)} = \frac{\tilde{\mathbf{b}}_i^{(k)}}{\lambda_i^{(k)}}$
  end
end

An efficient algorithm to solve MEP: Algorithm 2 is a Gauss-Seidel algorithm [34] to solve MEP which is proposed by [23]. This algorithm is essentially a variant of the classical power iteration method (see e.g. [38]). Let

$$r(\mathbf{b}^{(k)}) = \left(\mathbf{b}^{(k)}\right)^T C\mathbf{b}^{(k)},$$

$$\lambda_i(\mathbf{b}) = \mathbf{b}_i^T[C_{i1}, \ldots, C_{in}]\mathbf{b},$$

and

$$\Lambda(\mathbf{b}) = \mathrm{diag}(\lambda_1(\mathbf{b})I_{|\mathcal{X}_1|}, \ldots, \lambda_n(\mathbf{b})I_{|\mathcal{X}_n|}).$$

Theorem 7 ([26]) *Suppose the matrix C is symmetric. We have*

a) The sequence $\{r(\mathbf{b}^{(k)})\}$ generated by Algorithm 2 is monotonically increasing and convergent.

b) Let $(\Lambda^*, \mathbf{b}^*)$ be a solution of MEP. If $\mathbf{b}^*$ is a local maximizer of (5.7), then for any i, we have $\lambda_i(\mathbf{b}^*) \geq \sigma_{|\mathcal{X}_i|}(C_{ii})$. Moreover, if $\mathbf{b}^*$ is a global maximizer of (5.7), then for any i, we have $\lambda_i(\mathbf{b}^*) \geq \sigma_1(C_{ii})$, where

$$\sigma_1(C_{ii}) \geq \cdots \geq \sigma_{|\mathcal{X}_i|}(C_{ii})$$

are eigenvalues of the matrix C_{ii}.

A strategy for avoiding local optimums of MCP: Let $\mathbf{b}^*$ be a solution of MEP. Using Theorem 7, since C_{ii} is a zero matrix, in order to have $\mathbf{b}^*$ a global maximizer of optimization (5.7), we need to have $\lambda_i(\mathbf{b}^*) \geq 0$. Based on this observation, we have the following strategy for choosing an starting point for Algorithm 2. Let $\bar{\mathbf{b}}$ be a solution of (5.2) with the corresponding $\bar{\Lambda}$. Suppose that there exist an $1 \leq i \leq n$ such that $\lambda_i < 0$. Let $\mathbf{w}$ be the unit vector associated with the eigenvalue of $\bar{\lambda}_i I_{|\mathcal{X}_i|}$. Now let

$$\hat{\mathbf{b}} = \bar{\mathbf{b}} - \mathbf{q},$$

[2]Square root of a symmetric positive semidefinite matrix A is defined as $\sqrt{A} = U\Sigma^{1/2}U^T$ where $A = U\Sigma U^T$.

Algorithm 3 Network ACE to compute NMC

Input: G, $X_1, \ldots, X_n$,
Initialization: $\phi_1^{(0)}(X_1), \ldots, \phi_n^{(0)}(X_n)$ with mean zero.
for $k = 0, \ldots$
  $\phi_i(X_i) = \phi_i^{(k)}(X_i)$, $1 \le i \le n$
  for $i = 1 : n$
    $\phi_i^*(X_i) = \mathbb{E}\left[\sum_{j \in \mathcal{N}_i} \phi_j(X_j) | X_i\right]$
    update: $\phi_i(X_i) = \frac{\phi_i^*(X_i)}{\sqrt{\mathbb{E}[\phi_i^*(X_i)^2]}}$
  end
  $\phi_i^{(k+1)}(X_i) = \phi_i^*(X_i)$, $1 \le i \le n$
  $\rho_G^{(k+1)} = \sum_{(i,j) \in E} \mathbb{E}\left[\phi_i^{(k+1)}(X_i)\phi_j^{(k+1)}(X_j)\right]$
end

where $\mathbf{q}$ is a vector where $\mathbf{q}_{i'} = 0$ for all $i' \neq i$ and $\mathbf{q}_i = 2\mathbf{w}^T\bar{\mathbf{b}}_i\mathbf{w}$ still satisfies $\|\hat{\mathbf{b}}_i\| = 1$ for all $i = 1, \ldots, n$ and gives $r(\hat{\mathbf{b}}) = r(\bar{\mathbf{b}}) - 4\lambda_i(\mathbf{w}^T\bar{\mathbf{b}}_i)^2 > r(\bar{\mathbf{b}})$. We repeat this process until we have $\bar{\lambda}_i \ge 0$ for all $i = 1, \ldots, n$. Note that this is not a sufficient condition for the global maximizer of (5.7) and is only a necessary condition as Theorem 7 shows. After repeating this procedure, the condition given in Theorem 7 holds. We then call upon Algorithm 2 to produce yet a better solution for (5.7).

Based on Algorithm 2, we introduce an algorithm to compute NMC using alternating conditional expectation. We prove that the proposed algorithm converges to the local optimum of the NMC optimization (4.1). We then use a strategy explained in this section to jump out of local maximizers. At each iteration of the algorithm, we update transformation functions as follows: Suppose at iteration r, transformation functions are ϕ_j^r. If we fix all variables except the transformation function of node i, an optimal solution of ϕ_i^{r+1} can be written as the normalized conditional expectation of functions of its neighbors (see Proposition 3). In each update, the objective function of the NMC optimization increases or stays the same.

Proposition 6 *The sequence* $\{\rho_G^{(k)}\}_{k=0}^{\infty}$ *generated by Algorithm 3 converges to a local optimum solution of NMC optimization* (4.1).

Proof A proof is presented in Section 10.11. ∎

Figure 2 illustrates the convergence of the ACE algorithm to compute NMC of six jointly Gaussian variables connected over a complete graph (for more details, see Example 5).

Proposition 7 *The computational complexity of each iteration of Algorithm 3 is*

$$\mathcal{O}(nDd_{max} + |E|), \tag{5.8}$$

where d_{max} *is the maximum node degree and* $D = \max_i |\mathcal{X}_i|$.

Similarly to Algorithm 2 for solving MCP, Algorithm 3 finds a local optimum solution. Once the algorithm terminates, using Theorem 7 and the strategy provided in the previous section, if the convergence point does not satisfy the necessary conditions for a global optimum, we update the starting point of Algorithm 3 and run it again to reach a yet better solution.

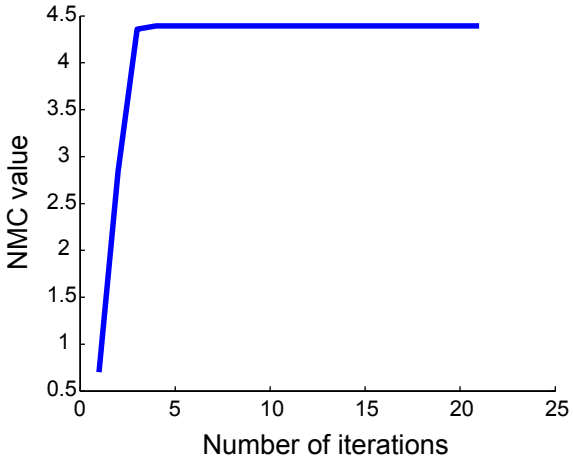

Figure 2: An illustration of the convergence of the ACE algorithm 3 to compute NMC over a complete graph connecting six jointly Gaussian variables.

Note that we can use a similar algorithm to Algorithm 3 to compute regularized NMC of Definition 5. The objective function of the regularized NMC optimization (4.15) can be written as follows:

$$\sum_{i\in V} \mathbb{E}[\phi_i(X_i)\ \big((1-\lambda)\sum_{j\in\mathcal{N}(i)}\phi_j(X_j)+\lambda X_i)] \tag{5.9}$$

Thus, to compute the regularized NMC, one can use a similar ACE Algorithm 3 with the following updates for transformation functions:

$$\phi_i^*(X_i) = \mathbb{E}[(1-\lambda)\sum_{j\in\mathcal{N}(i)}\phi_j(X_j)+\lambda X_i|X_i]. \tag{5.10}$$

If variables are continuous and we only observe samples from their joint distributions, empirical computation of conditional expectations in Algorithm 3 may be challenging owing to the lack of sufficient samples at each point. One way to overcome this issue is to discretize continuous variables by quantizing them. However, this approach can introduce significant quantization errors. An alternative approach to compute empirical conditional expectations at point $x_0 \in \mathbb{R}$ is to use all samples in its B neighborhood (i.e., $x_0 - B/2 \le x \le x_0 + B/2$). By using this approach, the computation of empirical conditional expectations in Algorithm 3 can be written as follows:

$$\phi_i^*(X_i = x_i) = \mathbb{E}[\sum_{j\in\mathcal{N}(i)}\phi_j(X_j)|x_i - B/2 \le X_i \le x_i + B/2]. \tag{5.11}$$

We use this approach in our ACE implementations to compute NMC for continuous variables. If the graph $G = (V, E)$ is sparse (i.e., $|E| = \mathcal{O}(n)$), Proposition 7 shows that the NMC computation can be performed efficiently in linear time complexity with respect to the number of nodes in the network. However, if the graph is dense or the number of nodes in the network is large, this computation may be expensive. In the following, we propose an approach to compute NMC using parallel computations.

5.3 Parallel Computation of NMC

For large and dense networks, exact computation of NMC may become computationally expensive (Proposition 7). For those cases, we propose a parallelizable algorithm which approximates NMC using network partitioning. The idea can be described as follow. For a given graph $G = (V, E)$,

(1) Partition the graph into small disjoint sets,

(2) Estimate NMC for each partition independently,

(3) Combine NMC solutions over sub-graphs to form an approximation of NMC for the original graph.

Definition 6 An (ϵ, k)- partitioning of graph $G = (V, E)$ is a distribution on finite partitions of V so that for any partition $\{V_1, \ldots, V_M\}$ with non zero probability, $|V_m| \geq k$, for all $1 \leq m \leq M$. Moreover, the probability that an edge falls across partitions is bounded by ϵ: for any $e \in E$, $\mathbb{P}[e \in E^c] \leq \epsilon$, where $E^c = E \setminus \bigcup_m (V_m \times V_m)$ is the set of cut edges. The probability is with respect to the distribution on partitions.

Definition 7 A graph G is poly-growth if there exists $r > 0$ and $C > 0$, such that for any vertex v in the graph,

$$|N_v(d)| \leq C d^r,$$

where $N_v(d)$ is the number of nodes within distance d of v in G.

Reference [39] describes the following procedure for generating an $(\epsilon, k)-$ partitioning on a graph:

1. Given $G = (V, E)$, k, and $\epsilon > 0$, we define the truncated geometric distribution as follows:

$$\mathbb{P}[x = l] = \begin{cases} \epsilon(1-\epsilon)^{l-1}, & l < k, \\ (1-\epsilon)^{k-1}, & l = k. \end{cases} \tag{5.12}$$

2. We then order nodes arbitrarily $1, \ldots, N$. For node i, we sample R_i according to distribution (5.12) and assign all nodes within that distance from node i to color i (distance is defined as the shortest path length on the graph). If a node has already colored, we re-color it with color i.

3. All nodes with the same color form a partition.

Proposition 8 *If G is a poly-graph, then by selecting $k = \Theta(\frac{r}{\epsilon} \log \frac{r}{\epsilon})$, the above procedure results in an (ϵ, Ck^r) partition.*

Proof See reference [39]. ∎

Next, we use an (ϵ, k)-graph partitioning to approximate NMC over large graphs using parallel computations. Consider the following approach:

(1) Given an (ϵ, k)- partitioning of G, we sample a partition $\{V_1, \ldots, V_M\}$ of V.

(2) For each partition $1 \le m \le M$, we compute NMC over $G_m = (V_m, E \cap (V_m \times V_m))$, denoted by $\hat{\rho}_{G_m}$.

(3) Let $\hat{\rho}_G = \sum_{m=1}^{M} \hat{\rho}(G_m)$ be an approximation of ρ_G.

In the following, we bound the approximation error by bounding boundary effects:

Theorem 8 *Consider an (ϵ, k)- partitioning of the graph G. We have,*

$$\mathbb{E}[\hat{\rho}_G] \ge (1-\epsilon)\rho_G, \tag{5.13}$$

where the expectation is over (ϵ, k)- partitions of graph G.

Proof A proof is presented in Section 10.12. ■

6 NMC Application in Inference of Nonlinear Gaussian Graphical Models

In this section, we discuss an application of NMC to infer graphical models for nonlinear functions of jointly Gaussian variables. Suppose $(X_1, \ldots, X_n)$ are jointly Gaussian variables with zero means and unit variances. Let $\rho_{i,i'}$ be the correlation coefficient between variables X_i and $X_{i'}$. Let $\psi_{i,j}$ to be the j-th Hermitte-Chebychev polynomial (2.8), for $1 \le i \le n$. Recall that these polynomials form an orthonormal basis with respect to Gaussian distribution (see Example 3 and reference [14] for convergence details). We have

$$\begin{aligned} \rho_{i,i'}^{j,j'} &= \mathbb{E}[\psi_{i,j}(X_i)\ \psi_{i',j'}(X_{i'})] \\ &= \rho_{i,i'}^{j} \mathbb{1}_{j=j'}, \end{aligned} \tag{6.1}$$

where $\mathbb{1}_{j=j'}$ is equal to one if $j = j'$, otherwise it is zero. Moreover, using the definition of Hermitte-Chebychev polynomials (2.8), we have

$$\mathbb{E}[\psi_{i,j}(X_i)] = \mathbb{1}_{j=0}, \quad 1 \le i \le n. \tag{6.2}$$

because all of these functions for $j \ge 1$ have zero means over a Gaussian distribution. Therefore, optimization (4.7) can be written as

$$\begin{aligned} \max \quad & \sum_{(i,i')\in E} \sum_{j=2}^{\infty} a_{i,j} a_{i',j}\ \rho_{i,i'}^{j} \\ & \sum_{j=2}^{\infty} (a_{i,j})^2 = 1, \quad 1 \le i \le n. \end{aligned} \tag{6.3}$$

In general, solving optimization (6.3) is NP-complete. We establish this by identifying that one instance of this optimization is simplified to the max-cut problem which is NP-complete [27].

Theorem 9 Let $s_i \in \{-1, 1\}$ for $1 \le i \le n$. Let $G = (V, E)$ be a complete graph. Suppose

$$\begin{aligned} &\sum_{i' \ne i} (1 - s_i s_{i'}) \rho_{i,i'} \ge 0, \quad \forall 1 \le i \le n, \\ &\sum_{i' \ne i} s_i s_{i'} \rho_{i,i'} \ge \sum_{i' \ne i} \rho_{i,i'}^2, \quad \forall 1 \le i \le n. \end{aligned}$$

Then, $\mathbf{a}_i^* = (0, s_i, 0, \ldots, 0)$, for $1 \le i \le n$ is a global maximizer of optimization (6.3).

Proof A proof is presented in Section 10.13. ■

Proposition 9 Under assumptions of Theorem 9, the NMC optimization (4.1) is simplified to the following max-cut optimization:

$$\begin{aligned} \max \quad & \sum_{i \neq i'} s_i s_{i'} \ \rho_{i,i'} \\ & s_i \in \{-1, 1\}, \quad 1 \leq i \leq n. \end{aligned} \tag{6.4}$$

Moreover, for all $1 \leq i \leq n$, we have $\phi_i^*(X_i) = s_i^* X_i$, where ϕ_i^* and s_i^* are optimal solutions of optimizations (3) and (6.4), respectively.

Proof A proof is presented in Section 10.14. ■

Note 2 Under the conditions of Theorem 9, one can compute the strength of the nonlinear relationships among variables by solving multiple pairwise MC optimization (2.1). However, one needs to solve optimization (6.4) to compute the signs of covariance coefficients.

In general, Max-Cut optimization (6.4) is NP-complete [27]. However, there exist algorithms to approximate its solution using Semidefinte Programming (SDP) with approximation factor of 0.87856 [28].

Corollary 1 Let $\phi_i^*(X_i)$ be an optimal solution of NMC optimization (3). Under assumptions of Theorem 9, if

$$\sum_{i' \neq i} \rho_{i,i'} \geq \sum_{i' \neq i} \rho_{i,i'}^2, \quad \forall 1 \leq i \leq n, \tag{6.5}$$

then, $\phi_i^*(X_i) = X_i$.

Intuitively, the assumptions of Corollary 1 considers jointly Gaussian variables with correlation coefficients that are mostly positive. However, the covariance matrix can have negative values as well. We will show that this assumption is critical in graphical model inference of nonlinear jointly Gaussian variables. Suppose $(X_1, \ldots, X_n)$ are jointly Gaussian variables with the covariance matrix Λ_X. Without loss of generality, we assume all variables have zero means and unit variances. I.e., $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^2] = 1$, for all $1 \leq i \leq n$. Let J_X be the information (precision) matrix [40] of these variables where $J_X = \Lambda_X^{-1}$. Define $G_X = (V_X, E_X)$ such that, $(i, j) \in E_X$ if and only if $J_X(i, j) \neq 0$.

Theorem 10 (e.g., [40]) *If* $(i, j) \notin E_X$, *then*

$$X_i \perp X_j | \{X_k, k \neq i, j\} \tag{6.6}$$

where $\perp$ *represents independency between variables.*

Theorem 10 represents a way to explicitly model the joint distribution of Gaussian variables using a graphical model $G_X = (V_X, E_X)$. This result is critical in several applications involved with Gaussian variables which requires computation of marginal distributions, or computation of the mode of the distribution. These computations can be performed efficiently over the graphical model using belief propagation approaches [40]. Moreover, Gaussian graphical models play an important

role in many applications such as linear regression [41], partial correlation [42], maximum likelihood estimation [43], etc. In many applications, even if variables are not jointly Gaussian, a Gaussian approximation is used often in practice, partially owing to the efficient inference of their graphical models.

In the following, under some conditions, we use the NMC optimization to characterize graphical models for functions of latent jointly Gaussian variables. These functions are unknown, bijective, and can be linear or nonlinear. More precisely, let $Y_i = f_i(X_i)$, where $f_i : \mathbb{R} \to \mathbb{R}$ is a bijective and differentiable function. Our goal is to characterize a graphical model for variables $(Y_1, Y_2, \ldots, Y_n)$ without the knowledge of $f_i(\cdot)$ functions. Consider the following NMC optimization:

$$\begin{aligned} \max \quad & \sum_{(i,i')} \mathbb{E}[g_i(Y_i)\ g_{i'}(Y_{i'})], \\ & \mathbb{E}[g_i(Y_i)] = 0, \quad 1 \le i \le n, \\ & \mathbb{E}[g_i^2(Y_i)] = 1, \quad 1 \le i \le n. \end{aligned} \tag{6.7}$$

Suppose $g_i^*(\cdot)$ represents an optimal solution for optimization (6.7). Define the matrix Λ_{nmc} such that $\Lambda_{nmc}(i,j) = \mathbb{E}[g_i^*(Y_i)g_j^*(Y_j)]$. Moreover, let $J_{nmc} = \Lambda_{nmc}^{-1}$. Define $G_{nmc} = (V_{nmc}, E_{nmc})$ such that, $(i,j) \in E_{nmc}$ iff $J_{nmc}(i,j) \neq 0$. The following theorem characterizes the graphical model of variables $(Y_1, \ldots, Y_n)$.

Theorem 11 *Suppose X_i's satisfy the conditions of Corollary 1. If $(i,j) \notin E_{nmc}$, then*

$$Y_i \perp Y_j | \{Y_k, k \neq i, j\}. \tag{6.8}$$

Proof A proof is presented in Section 10.15. ■

Under the conditions of Corollary 1, Theorem 11 characterizes the graphical model of variables Y_i's that are related to latent jointly Gaussian variables X_i's through unknown bijective functions f_i's. The family of distributions considered in Theorem 11 is broad and includes many Gaussian distributions as well as distributions whose variables are bijective functions of Gaussian variables. Graphical models characterized in Theorem 11 can be used in computation of marginal distributions, computation of the mode of the joint distribution, and in other applications of estimation and prediction similarly to the case of Gaussian graphical models. Note that Theorem 11 only considers f_i functions that are bijective and differentiable. If these functions were not bijective, the feasible set of optimization (6.7) is in fact smaller than the feasible set of the original NMC optimization (4.1) over Gaussian variables.

Example 5 Consider six jointly Gaussian variables X_1,...,X_6, each with unit variance and mean

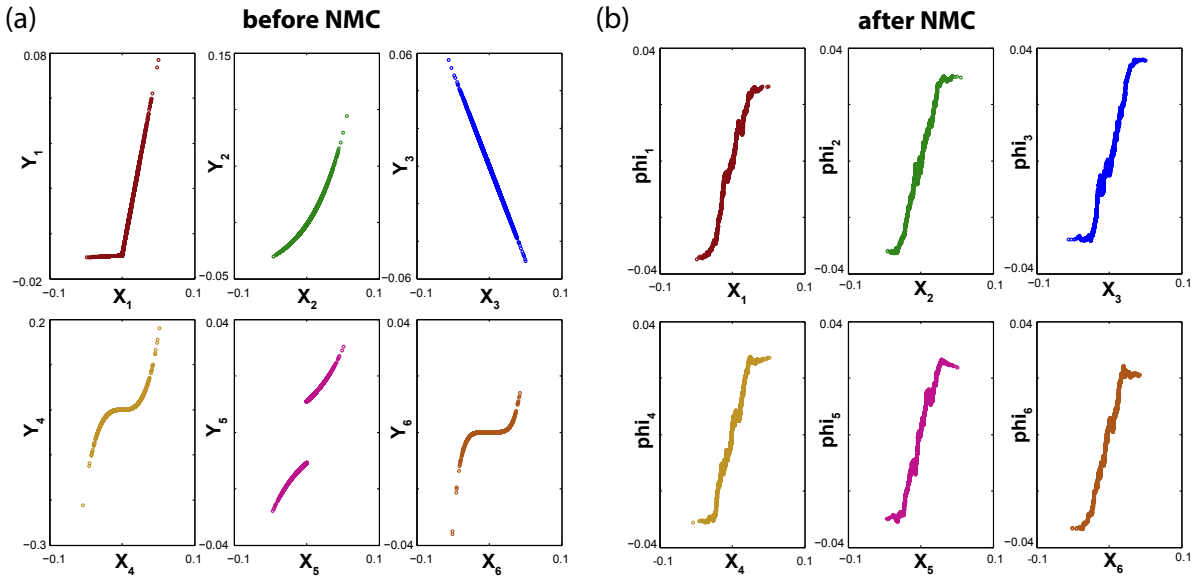

Figure 3: (a) Relationships among latent jointly Gaussian variables X_i and their nonlinear observations Y_i. (b) Relationships among latent jointly Gaussian variables and inferred transformations using the NMC optimization.

zero. We observe $Y_i = f_i(X_i)$ where,

$$
\begin{aligned}
Y_1 = f_1(X_1) &= \begin{cases} 10X_1, & \text{if } X_1 \geq 0, \\ \frac{1}{10}X_1, & \text{otherwise,} \end{cases} \\
Y_2 = f_2(X_2) &= e^{20X_2}, \\
Y_3 = f_3(X_3) &= -X_3, \\
Y_4 = f_4(X_4) &= X_4^3, \\
Y_5 = f_5(X_5) &= \begin{cases} e^{20X_5}, & \text{if } X_5 \geq 0, \\ -e^{-20X_5}, & \text{otherwise,} \end{cases} \\
Y_6 = f_6(X_6) &= X_6^5.
\end{aligned}
\tag{6.9}
$$

The functions $f_i(\cdot)$ remain unknown for the inference part. Relationships among original, observed and NMC variables are depicted in Figure 3. Then, we use $\phi_i^*(Y_i)$ to infer the underlying covariance and precision matrices according to Theorem 11. As it is illustrated in Figure 4, covariance and precision matrices computed using observed variables Y_i show significant errors compared to the true networks owing to the existence of extensive nonlinear relationships. However, inferred covariance and precision matrices using the NMC optimization closely approximate the true covariance and precision matrices, respectively (Theorem 11). Small errors in covariance coefficient estimation in this example are owing to computation of the NMC solution using empirical distributions according to the ACE algorithm 3.

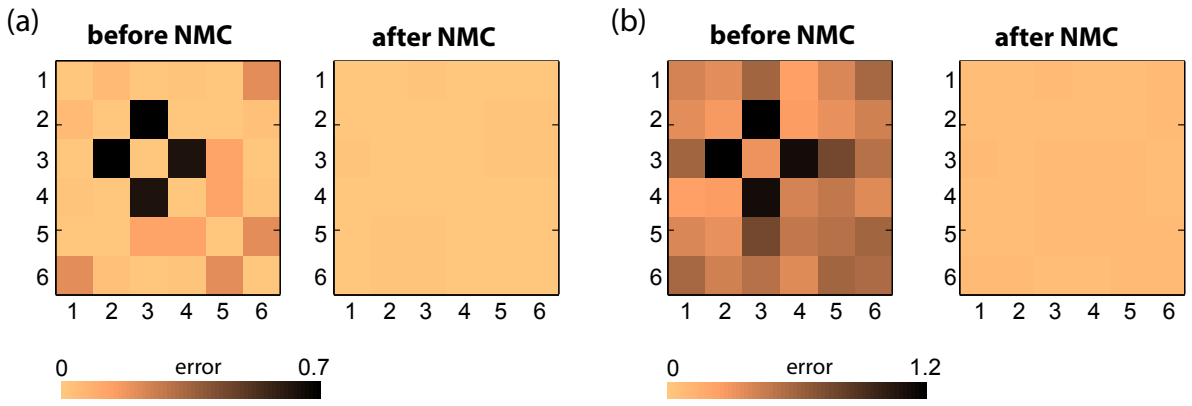

Figure 4: For six nonlinear functions of jointly Gaussian variables described in Example 5, panel (a) plots covariance matrix errors before and after NMC. Using NMC, estimated covariance coefficients among variables are close to true ones. Panel (b) plots precision matrix errors before and after NMC. Using NMC, estimated elements of the precision matrix are close to true ones.

7 NMC Application in Inference of Nonlinear Relevance Graphs

Relevance graphs (RG's) play an important role in many applications including systems biology, social and economic sciences as they characterize variable pairs with highest observed similarities [29]. Let X_1,...,X_n be n random variables with zero means and unit variances. Consider a similarity measure $S(X_i, X_j)$ between variables X_i and X_j. If the similarity measure between variables X_i and X_j is independent of other variables, the resulting graph is called a *pairwise relevance graph* (PRG).

Definition 8 Consider the following optimization:

$$G_S^* = \arg \max_{\substack{G=(V,E) \\ |E|=k}} \sum_{(i,j)\in E} S(X_i, X_j). \tag{7.1}$$

Then, $G_S^* = (V, E^*)$ is called a pairwise relevance graph (PRG) of variables X_1, ..., X_n, with k edges, corresponding to the similarity measure $S(\cdot, \cdot)$.

PRG's can be inferred efficiently in practice. In the following, we highlight two examples of such graphs:

Example 6 If $S(\cdot, \cdot)$ is a correlation-based similarity measure (e.g., $S(X_i, X_j) = |\mathbb{E}[X_i X_j]|$), the optimization (7.1) results in a correlation-based PRG. This graph only captures the top k linear associations among variables, ignoring nonlinear ones.

Example 7 If $S(\cdot, \cdot)$ is a mutual information-based similarity measure (i.e., $S(X_i, X_j) = I(X_i; X_j)$ where $I(.;.)$ represents the mutual information function [5]), optimization (7.1) results in an MI-based PRG which captures nonlinear associations among pairs of variables. However, it does not provide explicit forms of such nonlinear relationships.

Algorithm 4 Inference of a global relevance graph using NMC

Input: $X_1, \ldots, X_n$
Initialization: $E = \{\}$, $\mathcal{F}_1^* = \mathcal{F}_1, \ldots, \mathcal{F}_n^* = \mathcal{F}_n$
for $r = 1 : k$ **do**
  let $\{i, j, \phi_i^*, \phi_j^*\} = \text{argmax}\mathbb{E}[\phi_{i'}(X_{i'})\phi_{j'}(X_{j'})]$
  subject to: $\{i', j'\} \in G \times G \setminus E$, $\phi_{i'} \in \mathcal{F}_{i'}^*$, $\phi_{j'} \in \mathcal{F}_{j'}^*$
  update graph: $E = E \cup \{(i, j)\}$
  update: run Algorithm 3 on G
  For $i \in E$, let ϕ_i^* be the output function
  update function sets: For $i \in E$ let $\mathcal{F}_i^* = \{\phi_i^*\}$
end for
run Algorithm 3 on G to obtain $\rho_{G^*}(X_1, \ldots, X_n)$
Output: $E, \rho_{G^*}(X_1, \ldots, X_n), \phi_i^*(X_i), \ldots, \phi_n^*(X_n)$

If the similarity measure depends on all variables, the resulting graph is called a *global relevance graph* (GRG). A global relevance graph can capture system level properties of observed dependencies among variables. However, inference of a GRG is computationally challenging in general. Below, we introduce an NMC-based global relevance graph that captures observed nonlinear associations among variables and also provides explicit nonlinear transformation functions through which variables are associated.

Definition 9 The NMC-based GRG of variables X_1, ..., X_n, with k edges is defined as

$$G_{NMC}^* = \arg \max_{\substack{G=(V,E) \\ |E|=k}} \rho_G(X_1, \ldots, X_n), \tag{7.2}$$

where ρ_G is defined in (4.1).

Optimization (7.2) is combinatorial since it requires computation of NMC over all graphs with k edges. However, unlike MI-based GRG's, the required sample size to have a reliable computation of G_{NMC}^* does not increase with the size of the network. Below, we propose a greedy algorithm to find an approximate solution for optimization (7.2). Suppose at iteration r, the inferred network is $G = (V, E)$ whose nodes are assigned to transformations ϕ_i^r. At this iteration, for each non-connected pair of nodes, we wish to add the corresponding edge to the network, compute NMC, and finally select the edge with the highest NMC increase. However, this is computationally expensive as it requires NMC computation multiple times. Instead, at this iteration, we add an edge to the network with the highest correlation of its node transformations inferred in the last iteration. Then, we update all transformation functions by applying the ACE algorithm 3. We repeat this procedure explained in Algorithm 4 till the inferred network has a certain number of edges. Algorithm 4 starts from a graph with no edge and gradually populates it until the inferred graph has k edges. Alternatively, one can start with a complete graph and remove interactions with lowest nonlinear correlations iteratively. Because NMC is more efficient to run over sparse graphs (Proposition 7), the former has lower computational complexity than the later.

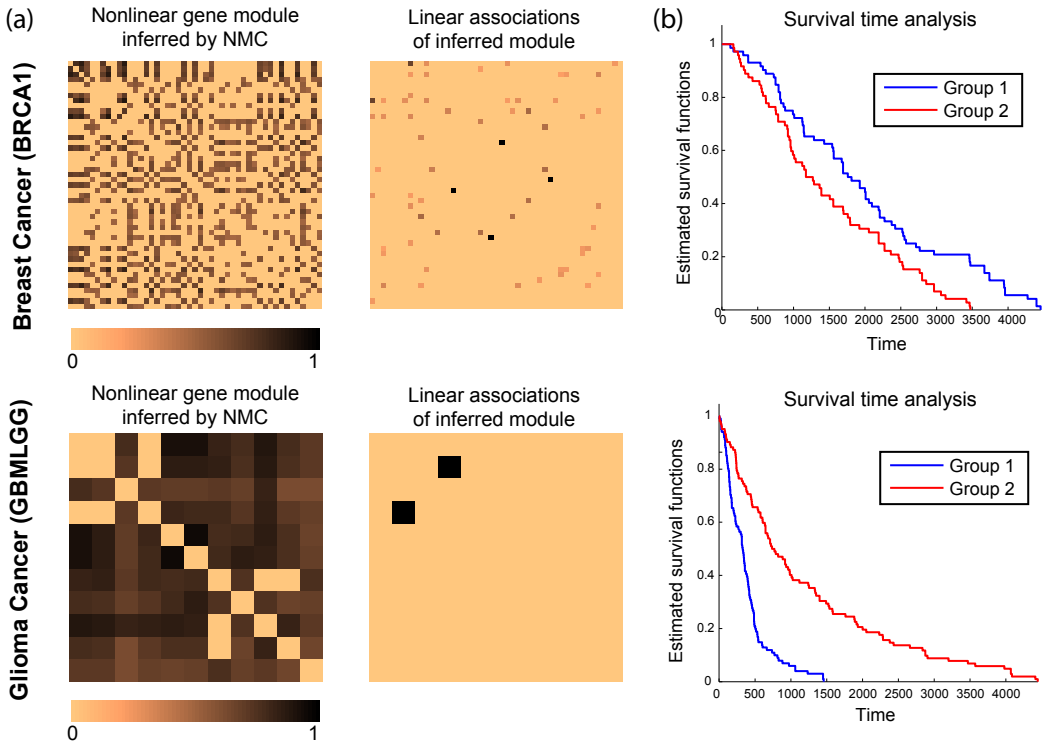

Figure 5: (a) Examples of nonlinear gene modules defined in Section 8, as a group of genes that is enriched in the NMC network but not in the linear one. (b) Survival time plots for corresponding nonlinear cancer modules of panel (a). For each inferred nonlinear cancer module, we partition individuals to two equal-size groups based on their average gene expression ranks in that module. We then perform a standard survival time analysis for each module and compute its associated log-rank p-value to determine its association with individual survival times in the considered cancer type.

8 Inference of Nonlinear Gene Modules in Cancer

Cancer is a complex disease involving abnormal cell growth with the potential to invade or spread to other parts of the body [44]. Different studies have shown associations of micro RNA patterns in different human cancers [30, 45]. In this section, we use NMC algorithms to infer gene modules of different cancer types that are detected over the nonlinear association network and not in the linear one. To perform these experiments, we use normalized RNA sequence counts from TCGA data portal for 24 cancer types at the gene level. We use processed the data provided in reference [30]. For each cancer type, first we select top 500 highly-variant genes based on their normalized variances [46]. Then, for each cancer, we compute both linear and nonlinear associations among genes. To compute NMC, we assume that conditions of Theorem 9 holds. That is, we assume that input data comes from, possibly nonlinear, functions of some latent jointly Gaussian variables satisfying conditions of Theorem 9. These functions are unknown and bijective. In general, this assumption is less restrictive than the assumption that input variables are jointly Gaussian.
In this application, we wish to infer the strength of associations among genes in different cancers.

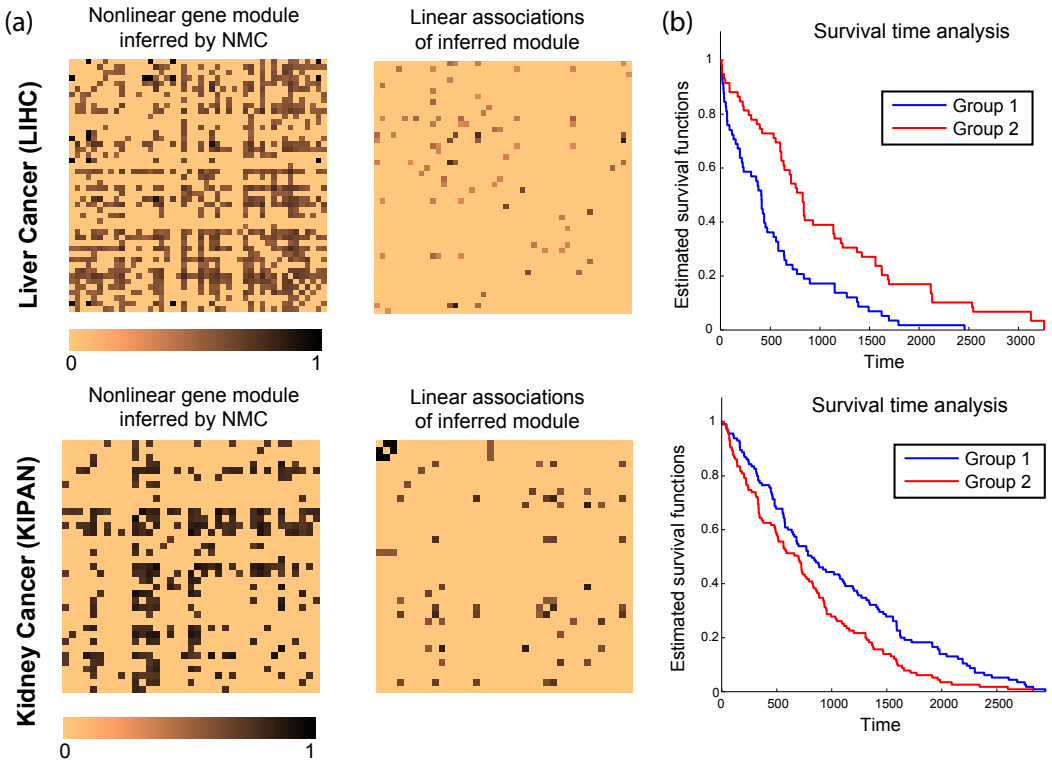

Figure 6: (a) Examples of nonlinear gene modules defined in Section 8, as a group of genes that is enriched in the NMC network but not in the linear one. (b) Survival time plots for corresponding nonlinear cancer modules of panel (a). For each inferred nonlinear cancer module, we partition individuals to two equal-size groups based on their average gene expression ranks in that module. We then perform a standard survival time analysis for each module and compute its associated log-rank p-value to determine its association with individual survival times in the considered cancer type.

To compute an NMC association network, according to Note 2, we compute multiple MC's among gene pairs using Algorithm 1. Moreover, we infer a linear association network by computing all pairwise correlations among gene expressions. We then select 5% of interactions among genes in the NMC network with largest nonlinear association increases compared to their linear association strengths. To have the same linear and nonlinear network densities, we select top 5% of interactions in the linear association network as well.

Next, we partition each network to k groups using a standard spectral clustering algorithm based on the modularity transformation [47]. We use $k = 10$ in all cases as it leads to dense and large clusters. We define *a gene module* as a group of genes that are densely connected to each other in the network. We compute a p-value for each gene module in the network by permuting the network structure and comparing the density of the module in the original network with the ones in permutated networks. We only consider gene modules with p-values less than 0.05. A gene module is called nonlinear if it is present in the NMC network but not in the linear one. Figures 5 and 6 demonstrate examples of inferred nonlinear gene modules in different cancer types.

For each inferred nonlinear module in a cancer type, we partition individuals to two equal-size

groups based on their average gene expression ranks in that module. We then perform a standard survival time analysis for each module based on Kaplan-Meier procedure to estimate survival function [48], and compute its associated log-rank p-value to determine its association with individual survival times in the considered cancer type [49]. We perform Benjamini and Hochberg multiple hypothesis correction [50] over the computed p-values of different nonlinear modules. For nonlinear modules with corrected p-values less than 0.05, we do further permutation analysis as follows: we randomly select the same number of genes as in the considered module, and compute its associated survival time p-value. If the corrected p-value of the nonlinear module is less than p-values of permutated modules at least in 95% of cases, we declare that nonlinear module as significantly associated with individual survival times in that cancer type.

Figures 5 and 6 illustrate examples of nonlinear gene modules for Breast Cancer (BRCA1), Glioma Cancer (GBMLGG), Liver Cancer (LIHC), and Kidney Cancer (KIPAN). These gene modules are significantly associated with survival times of individuals while they are not detected using linear association measures. Several references [51–55] have hypothesized that complex nonlinear relationships among genes may play important roles in cancer pathways. Our proposed NMC algorithms and inferred nonlinear gene modules can be used in discovering such complex nonlinear relationships in different cancer types. However, further experiments should be performed to determine the involvement of these nonlinear gene modules in different cancers, which is beyond the scope of this paper.

9 Conclusion

In this paper, we proposed Network Maximal Correlation (NMC) as a fundamental measure to capture nonlinear associations over networks without the knowledge of underlying nonlinearity shapes. We showed that NMC extends the standard Maximal Correlation to the case of having large number of variables, by assigning each variable to a single transformation function, thus avoiding over-fitting issues of using multiple MC optimizations over network edges. We also introduced a regularized NMC optimization which penalizes total distances of inferred nonlinear transformations from original variables. One can use other standard regularization techniques to further restrict inferred nonlinear functions in practical applications.

One of the main contributions of this work is providing a unifying framework to compute NMC (and therefore, MC) for both discrete and continuous variables using projections over appropriate Hilbert spaces. Using this framework, we established a connection between the NMC optimization with the MCP and MEP for discrete random variables, and with the Max-Cut problem for jointly Gaussian variables. Using these relationships, we provided efficient algorithms to compute NMC in different cases. To compute NMC for continuous random variables with general distributions, one can use the proposed optimization framework by choosing appropriate orthonormal basis for Hilbert spaces. For example, we used projections over Hermitte-Chebychev polynomials to characterize an optimal solution of the NMC optimization for jointly Gaussian variables.

Compared to other nonlinear association measures such as mutual information (MI), NMC has two main advantageous: first, unlike MI-based measures that only compute the strength of associations among variables, NMC characterizes nonlinear transformation functions through which variables are related to each other. These inferred, possibly nonlinear, functions can be used in different applications such as regression. As an example, a nonlinear regression framework [56] can be efficiently designed using transformations of variables. Second, for the case of having large number

of variables, a reliable computation of conditional MI requires an excessive number of samples which often is not available in practice. Here, we showed that NMC can be reliably computed in practice and provided a finite sample generalization bound and error guarantees.
NMC can be used in different areas to characterize nonlinear relationships, modules, and pathways among variables. Here, as an example, we applied NMC to different cancer datasets including breast, kidney and liver cancers, and showed that using NMC networks, we can infer nonlinear gene modules that are not detectable using linear association measures while they are significantly associated with survival times of individuals. Similarly, NMC can be applied to infer nonlinear gene interactions, modules and pathways in different types of biological networks such as regulatory [57] and PPI [58] networks. Moreover, NMC can be used over social/economic networks to characterize belief/behaviour variations of individuals/entities through their interactions over the underlying networks [59]. We believe that the proposed NMC framework and algorithms can make a significant impact in many areas of network sciences, statistics, information theory, systems biology, social sciences, and beyond.

10 Proofs

10.1 Proof of Proposition 1

Let $\{\psi_{1,i}\}_{i=1}^{\infty}$ and $\{\psi_{2,i}\}_{i=1}^{\infty}$ be the corresponding orthonormal bases of H_1 and H_2 in Definition 2. We can represent functions $\phi_1(X_1)$ and $\phi_2(X_2)$ in terms of the basis functions as follows:

$$\phi_1(x_1) = \sum_{i=1}^{\infty} a_{1,i}\psi_{1,i}(x_1),$$
$$\phi_2(x_2) = \sum_{i=1}^{\infty} a_{2,i}\psi_{2,i}(x_2),$$

for two sequence of coefficients $\{a_{1,i}\}_{i=1}^{\infty}$ and $\{a_{2,i}\}_{i=1}^{\infty}$. Thus, the constraint $\mathbb{E}[\phi_i(X_i)^2] = 1$ in optimization (2.1) would be translated into $\sum_{j=1}^{\infty} a_{i,j}^2 = 1$ and the constraint $\mathbb{E}[\phi_i(X_i)] = 0$ is simplified to $\sum_{j=1}^{\infty} a_{i,j}\mathbb{E}[\psi_{i,j}(X_i)] = 0$, for $i = 1, 2$.
Moreover, we have

$$\mathbb{E}[\phi_1(X_1)\phi_2(X_2)] = \sum_{i,j=1}^{\infty} a_{1,i}a_{2,j}\ \mathbb{E}[\psi_{1,i}(X_1)\psi_{2,j}(X_2)]. \tag{10.1}$$

Thus, maximal correlation optimization (2.1) can be re-written as follows:

$$\begin{aligned} \min \quad & \sum_{i,j} a_{1,i}a_{2,j}\ \mathbb{E}[\psi_{1,i}(X_1)\psi_{2,j}(X_2)] \\ & \sum_{j=1}^{\infty} a_{i,j}^2 = 1, \quad i = 1, 2, \\ & \sum_{j=1}^{\infty} a_{i,j}\ \mathbb{E}[\psi_{i,j}(X_i)] = 0, \quad i = 1, 2. \end{aligned} \tag{10.2}$$

Moreover, $\{\psi_{1,i}\psi_{2,j}\}_{i,j}$ forms a basis for co-variate L_2 functions of X_1 and X_2 including the joint distribution function $P_{X_1X_2}(\cdot,\cdot)$. Therefore, we have

$$P_{X_1X_2}(x_1, x_2) = \sum_{i,j} \rho_{ij}\psi_{1,i}(x_1)\psi_{2,j}(x_2). \tag{10.3}$$

Using equation (10.3), we have

$$\rho_{ij} = \mathbb{E}[\psi_{1,i}(X_1)\psi_{2,j}(X_2)]. \tag{10.4}$$

Using (10.4) in optimization (10.2) leads to optimization (2.2). This completes the proof.

10.2 Proof of Theorem 1

We use the following lemma in the proof of Theorem 1.

Lemma 2 *(a) Let P and $\tilde{P}$ be the matrix form of two joint probability distribution on $(\mathcal{X}_1, \mathcal{X}_2)$, such that $P_{X_1,X_2}(x_1, x_2) = [P]_{x_1,x_2}$ and $\tilde{P}_{X_1,X_2}(x_1, x_2) = [\tilde{P}]_{x_1,x_2}$. We bound the difference between Q and $\tilde{Q}$, by the difference between P and $\tilde{P}$, as follows:*

$$\begin{aligned}
&\|Q - \tilde{Q}\|_2 = \|D_{X_1}(P)^{-\frac{1}{2}} P D_{X_2}(P)^{-\frac{1}{2}} - D_{X_1}(\tilde{P})^{-\frac{1}{2}} \tilde{P} D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 \leq \\
&\frac{1}{\sqrt{\delta_{X_1}(P)}} \|P\|_2 \|D_{X_2}(P)^{-\frac{1}{2}} - D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 + \frac{1}{\sqrt{\delta_{X_1}(P)}} \frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}} \|P - \tilde{P}\|_2 \\
&+ \frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}} \|\tilde{P}\|_2 \|D_{X_1}(P)^{-\frac{1}{2}} - D_{X_1}(\tilde{P})^{-\frac{1}{2}}\|_2.
\end{aligned} \tag{10.5}$$

(b) Furthermore, we have

$$\begin{aligned}
\|Q - \tilde{Q}\|_2 \leq &\frac{1}{\sqrt{\delta_{X_1}(P)}} \sqrt{D} \|D_{X_2}(P) - D_{X_2}(\tilde{P})\|_\infty \frac{1}{2\delta_{X_2}(P, \tilde{P})^3} + \frac{1}{\sqrt{\delta_{X_1}(P)}} \frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}} \sqrt{D} \|P - \tilde{P}\|_\infty \\
&+ \frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}} \sqrt{D} \|D_{X_1}(P) - D_{X_1}(\tilde{P})\|_\infty \frac{1}{2\delta_{X_2}(P, \tilde{P})^{3/2}},
\end{aligned} \tag{10.6}$$

where $D = \max\{|\mathcal{X}_1|, |\mathcal{X}_2|\}$, $\delta_{X_2}(P, \tilde{P}) = \min\{\delta_{X_2}(P), \delta_{X_2}(\tilde{P})\}$, and $\delta_{X_1}(P, \tilde{P}) = \min\{\delta_{X_1}(P), \delta_{X_1}(\tilde{P})\}$.

Proof We have

$$\begin{aligned}
&\|Q - \tilde{Q}\|_2 \leq \|D_{X_1}(P)^{-\frac{1}{2}} P D_{X_2}(P)^{-\frac{1}{2}} - D_{X_1}(\tilde{P})^{-\frac{1}{2}} \tilde{P} D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 \leq \\
&\|D_{X_1}(P)^{-\frac{1}{2}} P D_{X_2}(P)^{-\frac{1}{2}} - D_{X_1}(P)^{-\frac{1}{2}} P D_{X_2}(\tilde{P})^{-\frac{1}{2}} \\
&+ D_{X_1}(P)^{-\frac{1}{2}} P D_{X_2}(\tilde{P})^{-\frac{1}{2}} - D_{X_1}(\tilde{P})^{-\frac{1}{2}} \tilde{P} D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 \leq \\
&\|D_{X_1}(P)^{-\frac{1}{2}} P\|_2 \|D_{X_2}(P)^{-\frac{1}{2}} - D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 + \|D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 \|D_{X_1}(P)^{-\frac{1}{2}} P - D_{X_1}(\tilde{P})^{-\frac{1}{2}} \tilde{P}\|_2 \leq \\
&\|D_{X_1}(P)^{-\frac{1}{2}} P\|_2 \|D_{X_2}(P)^{-\frac{1}{2}} - D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 + \\
&\|D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 \|D_{X_1}(P)^{-\frac{1}{2}} P - D_{X_1}(P)^{-\frac{1}{2}} \tilde{P} + D_{X_1}(P)^{-\frac{1}{2}} \tilde{P} - D_{X_1}(\tilde{P})^{-\frac{1}{2}} \tilde{P}\|_2 \leq \\
&\|D_{X_1}(P)^{-\frac{1}{2}} P\|_2 \|D_{X_2}(P)^{-\frac{1}{2}} - D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 + \|D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 \|D_{X_1}(P)^{-\frac{1}{2}}\|_2 \|P - \tilde{P}\|_2 + \\
&\|D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 \|\tilde{P}\|_2 \|D_{X_1}(P)^{-\frac{1}{2}} - D_{X_1}(\tilde{P})^{-\frac{1}{2}}\|_2.
\end{aligned} \tag{10.7}$$

For any distribution $P_{X_1,X_2}(\cdot, \cdot)$, we have

$$\|D_{X_1}(P)^{-\frac{1}{2}}\|_2 = \max_{X_1 \in \mathcal{X}_1} \frac{1}{\sqrt{P_{X_1}(X_1)}} = \frac{1}{\sqrt{\delta_{X_1}(P)}} \tag{10.8}$$

and

$$\|D_{X_2}(P)^{-\frac{1}{2}}\|_2 = \max_{X_2 \in \mathcal{X}_2} \frac{1}{\sqrt{P_{X_2}(X_2)}} = \frac{1}{\sqrt{\delta_{X_2}(P)}}. \tag{10.9}$$

By substituting (10.8) and (10.9) in (10.7), we obtain

$$\begin{aligned}\|Q - \tilde{Q}\|_2 \leq \frac{1}{\sqrt{\delta_{X_1}(P)}}\|P\|_2\|D_{X_2}(P)^{-\frac{1}{2}} - D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 + \frac{1}{\sqrt{\delta_{X_1}(P)}}\frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}}\|P - \tilde{P}\|_2 \\ + \frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}}\|\tilde{P}\|_2\|D_{X_1}(P)^{-\frac{1}{2}} - D_{X_1}(\tilde{P})^{-\frac{1}{2}}\|_2. \end{aligned} \tag{10.10}$$

This completes the first part of the lemma. Furthermore, for a joint probability distribution matrix $P(\cdot,\cdot)$, we have $\|P\|_\infty \leq 1$ and $\|P\|_1 \leq 1$. Moreover, we have

Lemma 3 *For a given $m \times n$ matrix, A, we have*

$$\|A\|_2 \leq \sqrt{m}\|A\|_\infty,$$

and

$$\|A\|_2 \leq \sqrt{n}\|A\|_1.$$

Therefore, using Lemma 3, we have $\|P\|_2 \leq \sqrt{|\mathcal{X}_2|}\|P\|_1 \leq \sqrt{|\mathcal{X}_2|}$ and $\|P\|_2 \leq \sqrt{|\mathcal{X}_1|}\|P\|_\infty \leq \sqrt{|\mathcal{X}_1|}$. Using this, we obtain the following:

$$\begin{aligned}\|Q - \tilde{Q}\|_2 \leq \frac{1}{\sqrt{\delta_{X_1}(P)}}\sqrt{D}\|D_{X_2}(P) - D_{X_2}(\tilde{P})\|_\infty \frac{1}{2\delta_{X_2}(P,\tilde{P})^3} + \frac{1}{\sqrt{\delta_{X_1}(P)}}\frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}}\sqrt{D}\|P - \tilde{P}\|_\infty \\ + \frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}}\sqrt{D}\|D_{X_1}(P) - D_{X_1}(\tilde{P})\|_1 \frac{1}{2\delta_{X_1}(P,\tilde{P})^{3/2}}, \end{aligned} \tag{10.11}$$

where $D = \max\{|\mathcal{X}_1|, |\mathcal{X}_2|\}$, $\delta_{X_2}(P,\tilde{P}) = \min\{\delta_{X_2}(P), \delta_{X_2}(\tilde{P})\}$, and $\delta_{X_1}(P,\tilde{P}) = \min\{\delta_{X_1}(P), \delta_{X_1}(\tilde{P})\}$. This concludes the proof. ■

Now we prove Theorem 1. Let Q and $\tilde{Q}$ be the corresponding normalized distribution matrices of P and $\tilde{P}$, respectively. Suppose the joint distribution on $(\mathcal{X}_1, \mathcal{X}_2)$ is according to P. Using Example 2, we have $\rho(X_1, X_2) = \sigma_2$, where σ_2 is the second largest singular value of the matrix Q. Similarly, maximal correlation under the joint distribution $\tilde{P}$ is $\tilde{\sigma}_2$, where $\tilde{\sigma}_2$ is the second largest singular value of the matrix $\tilde{Q}$. Moreover, we have

Theorem 12 *Let A_1 and A_2 be two matrices of the same size. We have*

$$|\sigma_i^{(1)} - \sigma_i^{(2)}| \leq \|A_1 - A_2\|_2, \tag{10.12}$$

where σ_i^1 and σ_i^2 are the i-th largest singular values of A_1 and A_2, respectively.

Proof See e.g. reference [60]. ■

Using Theorem 12 and part (a) of Lemma 2, we have

$$|\sigma_2 - \tilde{\sigma}_2| \leq \|Q - \tilde{Q}\|_2 \leq \frac{1}{\sqrt{\delta_{X_1}(P)}}\|P\|_2\|D_{X_2}(P)^{-\frac{1}{2}} - D_{X_2}(\tilde{P})^{-\frac{1}{2}}\|_2 + \frac{1}{\sqrt{\delta_{X_1}(P)}}\frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}}\|P - \tilde{P}\|_2$$
$$+ \frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}}\|\tilde{P}\|_2\|D_{X_1}(P)^{-\frac{1}{2}} - D_{X_1}(\tilde{P})^{-\frac{1}{2}}\|_2. \tag{10.13}$$

Furthermore, using part (b) of Lemma 2, we have

$$|\sigma_2 - \tilde{\sigma}_2| \leq \frac{1}{\sqrt{\delta_{X_1}(P)}}\sqrt{D}\|D_{X_2}(P) - D_{X_2}(\tilde{P})\|_\infty \frac{1}{2\delta_{X_2}(P,\tilde{P})^{3/2}} + \frac{1}{\sqrt{\delta_{X_1}(P)}}\frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}}\sqrt{D}\|P - \tilde{P}\|_\infty$$
$$+ \frac{1}{\sqrt{\delta_{X_2}(\tilde{P})}}\sqrt{D}\|D_{X_1}(P) - D_{X_1}(\tilde{P})\|_\infty \frac{1}{2\delta_{X_2}(P,\tilde{P})^{3/2}}. \tag{10.14}$$

Let $\delta = \min\{\delta_{X_1}(P), \delta_{X_1}(\tilde{P}), \delta_{X_2}(P), \delta_{X_2}(\tilde{P})\}$ and $D = \max\{|\mathcal{X}_1|, |\mathcal{X}_2|\}$. Since, $\|P - \tilde{P}\|_\infty \leq \epsilon$, we have

$$|\sigma_2 - \tilde{\sigma}_2| \leq \frac{\epsilon}{2\delta^2}D\sqrt{D} + \frac{\epsilon}{\delta}\sqrt{D} + \frac{\epsilon}{2\delta^2}D\sqrt{D} \leq 2\frac{\epsilon}{\delta^2}D\sqrt{D}. \tag{10.15}$$

This completes the proof.

10.3 Proof of Theorem 2

Using Lemma 2 and union bound, we have

$$\mathbb{P}\left[|\sigma_2(Q_m) - \sigma_2(Q)| > \epsilon\right] \leq$$
$$\mathbb{P}\left[\frac{1}{\delta_{X_1}(P)}\|D_{X_2}(P^{(m)})^{-\frac{1}{2}} - D_{X_2}(P)^{-\frac{1}{2}}\|_2\|P\|_2 > \frac{\epsilon}{3}\right]$$
$$+ \mathbb{P}\left[\frac{1}{\delta_{X_1}(P)}\frac{1}{\delta_{X_2}(P^{(m)})}\|P^{(m)} - P\|_2 > \frac{\epsilon}{3}\right]$$
$$+ \mathbb{P}\left[\frac{1}{\delta_{X_2}(P^{(m)})}\|D_{X_1}(P^{(m)})^{-\frac{1}{2}} - D_{X_1}(P)^{-\frac{1}{2}}\|_2\|P^{(m)}\|_2 > \frac{\epsilon}{3}\right]$$

$$\leq \mathbb{P}\left[\|D_{X_2}(P^{(m)})^{-\frac{1}{2}} - D_{X_2}(P)^{-\frac{1}{2}}\|_\infty > \frac{\epsilon}{3}\delta_{X_1}(P)\frac{1}{\sqrt{D}}\right]$$
$$+ \mathbb{P}\left[\|P^{(m)} - P\|_\infty > \delta_{X_1}(P)\delta_{X_2}(P^{(m)})\frac{\epsilon}{3}\frac{1}{\sqrt{D}}\right]$$
$$+ \mathbb{P}\left[\|D_{X_1}(P^{(m)})^{-\frac{1}{2}} - D_{X_1}(P)^{-\frac{1}{2}}\|_\infty > \delta_Y(P^{(m)})\frac{\epsilon}{3}\frac{1}{\sqrt{D}}\right] \tag{10.16}$$

where $D = \max\{|\mathcal{X}_1|, |\mathcal{X}_2|\}$. Moreover, we have

Theorem 13 *Let P be a probability distribution over alphabets $\mathcal{X}$. Also, let $P^{(m)}$ denote the empirical probability distribution of X, obtained from m i.i.d. samples, $\{x_i\}_{i=1}^m$, drawn according to P. We have*

$$\mathbb{P}\left[\|P^{(m)} - P\|_\infty > \epsilon\right] \leq 4e^{-m\epsilon^2}. \tag{10.17}$$

Proof See e.g. references [61, 62]. ■

Using Theorem 13, we have

$$\mathbb{P}\left[\|P^{(m)}(X_2) - P(X_2)\| \geq \frac{1}{2}\delta_{X_2}(P)\right] \leq 4e^{-m\left(\frac{1}{2}\delta_{X_2}(P)\right)^2}.$$

This occurs with probability less than $\eta/2$ if $m \geq \frac{1}{4}\frac{-\log \eta/2}{(\frac{1}{2}\delta)^2}$. Moreover, given $\|P^{(m)}(X_2) - P(X_2)\| \geq \frac{1}{2}\delta_{X_2}(P)$, we obtain $\delta_{X_2}(P^{(m)}) > \frac{1}{2}\delta_{X_2}(P)$. We also have

$$\begin{aligned}
&\mathbb{P}\left[|\sigma_2(Q_m) - \sigma_2(Q)| > \epsilon\right] \\
&= \mathbb{P}\left[|\sigma_2(Q_m) - \sigma_2(Q)| > \epsilon \mid \|P^{(m)}(X_2) - P(X_2)\| \geq \frac{1}{2}\delta_{X_2}(P)\right]\mathbb{P}\left[\|P^{(m)}(X_2) - P(X_2)\| \geq \frac{1}{2}\delta_{X_2}(P)\right] \\
&+ \mathbb{P}\left[|\sigma_2(Q_m) - \sigma_2(Q)| > \epsilon \mid \|P^{(m)}(X_2) - P(X_2)\| \leq \frac{1}{2}\delta_{X_2}(P)\right]\mathbb{P}\left[\|P^{(m)}(X_2) - P(X_2)\| \leq \frac{1}{2}\delta_Y(P)\right] \\
&\leq \mathbb{P}\left[|\sigma_2(Q_m) - \sigma_2(Q)| > \epsilon \mid \|P^{(m)}(X_2) - P(X_2)\| \leq \frac{1}{2}\delta_Y(P)\right] + \mathbb{P}\left[\|P^{(m)}(X_2) - P(X_2)\| \geq \frac{1}{2}\delta_{X_2}(P)\right].
\end{aligned}$$

Let $m \geq \frac{1}{4}\frac{-\log \frac{\eta}{2}}{(\frac{1}{2}\delta)^2}$ to make the second term less than or equal to $\frac{1}{2}\eta$. Next, we choose m sufficiently large so that the first term is no greater than $\frac{1}{2}\eta$. Using (10.16), we have

$$\begin{aligned}
&\mathbb{P}\left[|\sigma_2(Q_m) - \sigma_2(Q)| > \epsilon \mid \|P^{(m)}(X_2) - P(X_2)\| \leq \frac{1}{2}\delta_{X_2}(P)\right] \\
&\leq \mathbb{P}\left[\|D_{X_2}(P^{(m)}) - D_{X_2}(P)\|_\infty > \frac{\epsilon}{3}\left(2\delta_{X_2}(P)^3\right)\delta_{X_1}(P)\frac{1}{\sqrt{D}}\right] \\
&+ \mathbb{P}\left[\|P^{(m)} - P\|_\infty > \delta_{X_1}(P)\frac{1}{2}\delta_{X_2}(P)\frac{\epsilon}{3}\frac{1}{\sqrt{D}}\right] \\
&+ \mathbb{P}\left[\|D_{X_1}(P^{(m)}) - D_{X_1}(P)\|_\infty > \left(2\delta_{X_1}(P)^3\right)\frac{1}{2}\delta_{X_2}(P)\frac{\epsilon}{3}\frac{1}{\sqrt{D}}\right].
\end{aligned}$$

Now using Theorem 13, each term on the righthand side goes to zero exponentially fast. Moreover, we have $\mathbb{P}[|\sigma_2(Q_m) - \sigma_2(Q)| > \epsilon] \leq 12e^{-me^2}$, where

$$e = \min\{\frac{\epsilon}{3}\left(2\delta_{X_2}(P)^3\right)\delta_{X_1}(P)\frac{1}{\sqrt{D}}, \delta_X(P)\frac{1}{2}\delta_{X_2}(P)\frac{\epsilon}{3}\frac{1}{\sqrt{D}}, \left(2\delta_{X_1}(P)^3\right)\frac{1}{2}\delta_{X_2}(P)\frac{\epsilon}{3}\frac{1}{\sqrt{D}}\},$$

which is equal to $\frac{\epsilon}{3}\delta^4\frac{1}{\sqrt{D}}$. Therefore, in order to have $\mathbb{P}[|\sigma_2(Q_m) - \sigma_2(Q)| > \epsilon \mid \|P^{(m)}(X_2) - P(X_2)\| \leq \frac{1}{2}\delta_{X_2}(P)] \leq \frac{\eta}{2}$, it suffices to have

$$m \geq \frac{\log \frac{12}{\frac{\eta}{2}}}{\frac{\epsilon}{3}\delta^4\frac{1}{\sqrt{D}}}.$$

Therefore, overall it suffices to have $m \geq \frac{\log \frac{24}{\eta}}{\frac{\epsilon}{3}\delta^4\frac{1}{\sqrt{D}}}$. This concludes the proof.

10.4 Proof of Lemma 1

Consider the following MSE optimization. We have

$$\begin{aligned}&\min_{\phi_1,\ldots,\phi_n} \frac{1}{2}\sum_{(i,j)\in E}\mathbb{E}[(\phi_i(X_i)-\phi_j(X_j))^2]\\ =&|E|-\max_{\phi_1,\ldots,\phi_n}\sum_{(i,j)\in E}\mathbb{E}[\phi_i(X_i)\phi_j(X_j)].\end{aligned}$$

Therefore, the NMC optimization (4.1) is equivalent to the following MSE optimization:

$$\min_{\phi_1,\ldots,\phi_n} \frac{1}{2}\sum_{(i,j)\in E}\mathbb{E}[(\phi_i(X_i)-\phi_j(X_j))^2], \tag{10.18}$$

where $\mathbb{E}[\phi_i(X_i)]=0$ and $\mathbb{E}[\phi_i^2(X_i)]=1$, for any $1\le i\le n$.

10.5 Proof of Proposition 2

Let

$$F_i=\left\{\phi_i\in H_i:\quad \mathbb{E}[\phi_i]=0,\quad \mathbb{E}[\phi_i^2]=1\right\}.$$

Now consider the function defined on $F_1\times\cdots\times F_n$, as

$$R(\phi_1,\ldots,\phi_n)=\sum_{(i,j)\in E}\mathbb{E}[\phi_i(X_i)\phi_j(X_j)].$$

Below, we show that R is continuous and $F_1\times\cdots\times F_n$ is a compact subset of a finite dimensional space.

For any $1\le i\le n$, first we show F_i is compact. Since F_i is a subset of compact space H_i, it suffices to show that it is closed. This evidently follows from the definition of the norm on H_i, which is $\sqrt{\mathbb{E}[(\phi_i-\phi_i')^2]}$. Since $F_1\times\cdots\times F_n$ is the product of finitely many compact sets, therefore, it is compact as well.

Next we show R is continuous. Since R has finitely many arguments, it suffices to show that it is continuous with respect to each argument. Let $1\le i_0\le n$. We have

$$\begin{aligned}R(\phi_1,\ldots,\phi_n)&=\sum_{j\in\mathcal{N}(i_0)}\langle\phi_{i_0},\mathcal{P}_{i_0}\phi_j\rangle+\sum_{(i,j)\in E,\ i\neq i_0,\ j\neq i_0}\langle\phi_i,\phi_j\rangle\\&=\left\langle\phi_{i_0},\sum_{j\in\mathcal{N}(i_0)}\mathcal{P}_{i_0}\phi_j\right\rangle+\sum_{(i,j)\in E,\ i\neq i_0,\ j\neq i_0}\langle\phi_i,\phi_j\rangle,\end{aligned}$$

where $\mathcal{P}_i$ denotes the projection operation from the space H_j (for any $j\neq i$) onto H_i. Moreover, we can employ on Weierstrass Extreme Value Theorem that says every continuous real-valued function on a compact space attains its extreme values. Since $\left\langle\phi_{i_0},\sum_{j\in\mathcal{N}(i_0)}\mathcal{P}_{i_0}\phi_j\right\rangle$ is continuous with respect to ϕ_{i_0}, the function $R:F_1\times\cdots\times F_n\to\mathbb{R}$ is continuous. Thus, function R attains its maximum on $F_1\times\cdots\times F_n$, which complete the proof.

10.6 Proof of Proposition 3

For a given $1 \le i_0 \le n$, we have

$$\begin{aligned}\sum_{(i,j)\in E} \langle \phi_i, \phi_j \rangle &= \sum_{j\in\mathcal{N}(i_0)} \langle \phi_{i_0}, \mathcal{P}_{i_0}\phi_j \rangle + \sum_{(i,j)\in E,\ i\neq i_0,\ j\neq i_0} \langle \phi_i, \phi_j \rangle \\ &= \left\langle \phi_{i_0}, \sum_{j\in\mathcal{N}(i_0)} \mathcal{P}_{i_0}\phi_j \right\rangle + \sum_{(i,j)\in E,\ i\neq i_0,\ j\neq i_0} \langle \phi_i, \phi_j \rangle \\ &\le \| \sum_{j\in\mathcal{N}(i_0)} \mathcal{P}_{i_0}\phi_j \| + \sum_{(i,j)\in E,\ i\neq i_0,\ j\neq i_0} \langle \phi_i, \phi_j \rangle ,\end{aligned}$$

where the last inequality that follows from Cauchy-Schwartz inequality becomes an equality if and only if we have $\phi^*_{i_0} = c\sum_{j\in\mathcal{N}(i_0)} \mathcal{P}_{i_0}\phi^*_j$, for some constant $c > 0$. Since we have $\|\phi^*_{i_0}\| = 1$, we obtain

$$\phi^*_{i_0} = \frac{\sum_{j\in\mathcal{N}(i_0)} \mathcal{P}_{i_0}\phi^*_j}{\| \sum_{j\in\mathcal{N}(i_0)} \mathcal{P}_{i_0}\phi^*_j \|},$$

which completes the proof.

10.7 Proof of Theorem 3

Since any Hilbert space has an orthonormal basis (Theorem 2.4, [33]), let $\{\psi_{i,j}\}_{j=1}^{\infty}$ be the corresponding orthonormal basis of H_i for $1 \le i \le n$. Therefore, we can represent $\phi_i(X_i)$ in terms of the basis as follows:

$$\phi_i(x_i) = \sum_{j=1}^{\infty} a_{i,j}\psi_{i,j}(x_i).$$

The constraint $\mathbb{E}[\phi_i(X_i)^2] = 1$ translates into $\sum_{j=1}^{\infty} a_{i,j}^2 = 1$ and the constraint $\mathbb{E}[\phi_i(X_i)] = 0$ translates into $\sum_{j=1}^{\infty} a_{i,j}\mathbb{E}[\psi_{i,j}(X_i)] = 0$ for $1 \le i \le n$. Moreover, for any i and i' we have

$$\mathbb{E}[\phi_i(X_i)\phi_{i'}(X_{i'})] = \sum_{j,j'} a_{i,j}a_{i',j'}\mathbb{E}[\psi_{i,j}(X_i)\psi_{i',j'}(X_{i'})],$$

The network maximal correlation can be found by solving the following optimization problem.

$$\begin{aligned}\max \quad & \sum_{(i,i')\in E}\sum_{j,j'} a_{i,j}a_{i',j'}\ \mathbb{E}[\psi_{i,j}(X_i)\psi_{i',j'}(X_{i'})] \\ & \sum_{j=1}^{\infty} a_{i,j}^2 = 1, \quad 1 \le i \le n, \\ & \sum_{j=1}^{\infty} a_{i,j}\ \mathbb{E}[\psi_{i,j}(X_i)] = 0, \quad 1 \le i \le n,\end{aligned}$$

Moreover, since $\{\psi_{i,j}\psi_{i',j'}\}$ is a basis for functions of X_i and $X_{i'}$ (with the corresponding inner product), we can write the joint density function as

$$P_{X_iX_{i'}}(X_i, X_{i'}) = \sum_{j,j'} \rho_{i,i'}^{j,j'}\psi_{i,j}(X_i)\psi_{i',j'}(X_{i'}).$$

Using this equation, we have $\mathbb{E}[\psi_{i,j}(X_i)\psi_{i',j'}(X_{i'})] = \rho_{i,i'}^{j,j'}$, which completes the proof.

10.8 Proof of Theorem 4

Let P and $\tilde{P}$ be two distributions over $(X_1, \ldots, X_n)$. We shall compare the solution of the two following optimization problems.

$$\begin{aligned} \max_{\mathbf{a}_i} \quad & \sum_{(i,i')\in E} \mathbf{a}_i^T Q_{i,i'} \mathbf{a}_i \\ & \|\mathbf{a}_i\|_2 = 1, \quad 1 \le i \le n, \\ & \mathbf{a}_i \perp \sqrt{\mathbf{p}_i}, \quad 1 \le i \le n, \end{aligned} \tag{10.19}$$

and

$$\begin{aligned} \max_{\mathbf{a}_i} \quad & \sum_{(i,i')\in E} \mathbf{a}_i^T \tilde{Q}_{i,i'} \mathbf{a}_i \\ & \|\mathbf{a}_i\|_2 = 1, \quad 1 \le i \le n, \\ & \mathbf{a}_i \perp \sqrt{\tilde{\mathbf{p}}_i}, \quad 1 \le i \le n. \end{aligned} \tag{10.20}$$

Let ρ_G and $\tilde{\rho}_G$ be the optimal solution for (10.19) and (10.20), respectively. For any $(i,j) \in E$, suppose $\|P - \tilde{P}\|_\infty \le \epsilon$. Also, suppose $\delta = \min_{1\le i\le n} \left(\min\{\delta_{X_i}(P_{X_i}), \delta_{X_i}(\tilde{P}_{X_i})\}\right)$. Let the optimal solution of optimization (10.19) be $\mathbf{a}_i^*$. Based on this solution, we shall construct a feasible solution for optimization (10.20) and then evaluate its objective function.
For any i, let

$$\mathbf{b}_i = \frac{\left(\mathbf{a}_i^* + \sqrt{\tilde{\mathbf{p}}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} >\right)}{\|\mathbf{a}_i^* + \sqrt{\tilde{\mathbf{p}}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|}.$$

We claim this set of vectors is feasible for optimization (10.20). First note that the norm of each $\mathbf{b}_i$ is one. We next show that each $\mathbf{b}_i$ is orthogonal to $\sqrt{\tilde{\mathbf{p}}_i}$.

$$< \mathbf{b}_i, \sqrt{\tilde{\mathbf{p}}_i} >= \frac{1}{\|\mathbf{a}_i^* + \sqrt{\tilde{\mathbf{p}}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|} \left(< \mathbf{a}_i^*, \sqrt{\tilde{\mathbf{p}}_i} > + < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|\sqrt{\tilde{\mathbf{p}}_i}\| \right) = 0,$$

where the last equality follows from $\|\sqrt{\tilde{\mathbf{p}}_i}\| = 1$. We now plug in the feasible solution $\mathbf{b}_i$ into the objective function of optimization (10.20).

$$\begin{aligned} \tilde{\rho}_G &\ge \sum_{(i,i')\in E} \mathbf{b}_i^T \tilde{Q}_{ii'} \mathbf{b}_{i'} = \sum_{(i,i')\in E} \mathbf{b}_i^T \left(\tilde{Q}_{ii'} - Q_{ii'}\right) \mathbf{b}_{i'} + \left(\mathbf{b}_i^T - \mathbf{a}^{*T}_i\right) Q_{ii'} \mathbf{b}_{i'} + \mathbf{a}^{*T}_i Q_{ii'} \left(\mathbf{b}_{i'} - \mathbf{a}^*_{i'}\right) + \mathbf{a}^{*T}_i Q_{ii'} \mathbf{a}^*_{i'} \\ &= \rho_G + \sum_{(i,i')\in E} \mathbf{b}_i^T \left(\tilde{Q}_{ii'} - Q_{ii'}\right) \mathbf{b}_{i'} + \left(\mathbf{b}_i^T - \mathbf{a}^{*T}_i\right) Q_{ii'} \mathbf{b}_{i'} + \mathbf{a}^{*T}_i Q_{ii'} \left(\mathbf{b}_{i'} - \mathbf{a}^*_{i'}\right). \end{aligned} \tag{10.21}$$

We now bound each term on the right hand side of equation (10.21). Using Lemma 2, for any i, i', we have

$$|\mathbf{b}_i^T \left(\tilde{Q}_{ii'} - Q_{ii'}\right) \mathbf{b}_{i'}| \le \|\tilde{Q}_{ii'} - Q_{ii'}\|_2 \le \frac{3}{2} \frac{\epsilon}{\delta^2} D^{3/2}.$$

We also have

$$
\begin{aligned}
|\left(\mathbf{b}_i^T - \mathbf{a}^{*T}_i\right) Q_{ii'}\mathbf{b}_{i'}| &\leq \|\mathbf{b}_i^T - \mathbf{a}^{*T}_i\|_2 \|Q_{ii'}\|_2 \|\mathbf{b}_{i'}\|_2 = \|\mathbf{b}_i^T - \mathbf{a}^{*T}_i\|_2 \\
&\leq \frac{\|\sqrt{\tilde{\mathbf{p}}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|_2}{1 - \|\sqrt{\tilde{\mathbf{p}}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|_2} \\
&+ \max\{\frac{\|\sqrt{\tilde{\mathbf{p}}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|_2}{1 - \|\sqrt{\tilde{\mathbf{p}}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|_2}, \frac{\|\sqrt{\tilde{\mathbf{p}}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|_2}{1 + \|\sqrt{\tilde{\mathbf{p}}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|_2}\},
\end{aligned}
$$

where we use the following basic inequality

$$
\begin{aligned}
\|\frac{\mathbf{a}+\epsilon}{\|\mathbf{a}+\epsilon\|_2} - \mathbf{a}\|_2 &= \|\frac{\epsilon}{\|\mathbf{a}+\epsilon\|_2} + \mathbf{a}\left(\frac{1}{\|\mathbf{a}+\epsilon\|_2} - 1\right)\|_2 \\
&\leq \|\frac{\epsilon}{\|\mathbf{a}+\epsilon\|_2}\|_2 + \|\mathbf{a}\|_2 |\frac{1}{\|\mathbf{a}+\epsilon\|_2} - 1| \\
&\leq \frac{\|\epsilon\|_2}{1 - \|\epsilon\|_2} + \max\{|\frac{1}{1-\|\epsilon\|_2} - 1|, |\frac{1}{1+\|\epsilon\|_2} - 1|\},
\end{aligned}
$$

where we used $1 - \|\epsilon\|_2 \leq \|\mathbf{a}+\epsilon\|_2 \leq 1 + \|\epsilon\|_2$ to obtain the last inequality. Therefore, we have

$$
|\left(\mathbf{b}_i^T - \mathbf{a}^{*T}_i\right) Q_{ii'}\mathbf{b}_{i'}| \leq 2\frac{\|\sqrt{\tilde{p}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|_2}{1 - \|\sqrt{\tilde{p}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|_2}.
$$

We shall now bound the term $\|\sqrt{\tilde{\mathbf{p}}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|_2$. Using Cauchy-Schwartz, we have

$$
\|\sqrt{\tilde{\mathbf{p}}_i} < \mathbf{a}_i^*, \sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|_2 \leq \|\sqrt{\tilde{\mathbf{p}}_i}\|_2 \|\mathbf{a}_i^*\|_2 \|\sqrt{\mathbf{p}_i} - \sqrt{\tilde{\mathbf{p}}_i} > \|_2 \leq \frac{\epsilon}{2\delta^{1/2}} D^{3/2},
$$

where we use the fact that $|\mathbf{p}_i(j) - \tilde{\mathbf{p}}_i(j)| \leq \epsilon D$ and $|\sqrt{\tilde{\mathbf{p}}_i(j)} - \sqrt{\mathbf{p}_i(j)}| \leq |\mathbf{p}_i(j) - \tilde{\mathbf{p}}_i(j)| \frac{1}{2\delta^{1/2}}$. Plugging in this inequality and choosing ϵ sufficiently small to guarantee $\frac{\epsilon}{2\delta^{1/2}} D^{3/2} \leq \frac{1}{2}$, we obtain

$$
|\left(\mathbf{b}_i^T - \mathbf{a}^{*T}_i\right) Q_{ii'}\mathbf{b}_{i'}| \leq 2\frac{\epsilon}{\delta^{1/2}} D^{3/2},
$$

which leads to

$$
\begin{aligned}
\tilde{\rho}_G &\geq \rho_G - \sum_{(i,i')\epsilon E} 4\frac{\epsilon}{\delta^{1/2}} D^{3/2} + \frac{3\epsilon}{2\delta^2} D^{3/2} \\
&= \rho_G - \epsilon |E| D^{\frac{3}{2}} \left(\frac{4}{\delta^{1/2}} + \frac{3}{2\delta^2}\right).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
\rho_G &\geq \tilde{\rho}_G - \sum_{(i,i')\epsilon E} 4\frac{\epsilon}{\delta^{1/2}} D^{3/2} + \frac{3\epsilon}{2\delta^2} D^{3/2} \\
&= \rho_G - \epsilon |E| D^{\frac{3}{2}} \left(\frac{4}{\delta^{1/2}} + \frac{3}{2\delta^2}\right).
\end{aligned}
$$

Combining the previous two relations, we obtain

$$
|\tilde{\rho}_G - \rho_G| \leq \epsilon |E| D^{\frac{3}{2}} \left(\frac{4}{\delta^{1/2}} + \frac{3}{2\delta^2}\right) \leq \epsilon |E| D^{\frac{3}{2}} \frac{6}{\delta^2},
$$

which completes the proof.

10.9 Proof of Theorem 5

Using Theorem 13 for any $(i,i') \in E$, we have

$$\mathbb{P}[\|P_{i,i'}^{(m)} - P_{i,i'}\|_\infty > \gamma] \le 4e^{-m\gamma^2}.$$

Therefore, using the union bound, we obtain

$$\begin{aligned}\mathbb{P}[\cap_{(i,i')\in E}\mathbf{1}\{\|P_{i,i'}^{(m)} - P_{i,i'}\|_\infty < \gamma\}] &= 1 - \mathbb{P}[\cup_{(i,i')\in E}\mathbf{1}\{\|P_{i,i'}^{(m)} - P_{i,i'}\|_\infty > \gamma\}] \\ &\ge 1 - \sum_{(i,i')\in E}\mathbb{P}[\mathbf{1}\{\|P_{i,i'}^{(m)} - P_{i,i'}\|_\infty > \gamma\}] \ge 1 - 4|E|e^{-n\gamma^2}.\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\mathbb{P}[\delta(P_i^{(m)}) \ge \frac{\delta}{2}] &\ge \mathbb{P}[\|P_i^{(m)} - P_i\|_\infty \le \frac{\delta}{2}] \ge \mathbb{P}[|\|P_{i,i'}^{(m)}| - P_{i,i'}^{(m)}|_\infty \le \frac{\delta}{2D}] \\ &\ge 1 - \mathbb{P}[\|P_{i,i'}^{(m)} - P_{i,i'}^{(m)}\|_\infty > \frac{\delta}{2D}] \ge 1 - 4e^{-m(\frac{\delta}{2D})^2}.\end{aligned}$$

Therefore, using the union bound, we obtain

$$\mathbb{P}[\cap_{i\in V}\mathbf{1}\{\delta(P_i^{(m)}) \ge \frac{\delta}{2}\}] \ge 1 - 4|V|e^{-m(\frac{\delta}{2D})^2}.$$

By applying the union bound once more, we have

$$\mathbb{P}[\cap_{(i,i')\in E}\mathbf{1}\{\|P_{i,i'}^{(m)} - P_{i,i'}\|_\infty < \gamma\} \cap_{i\in V} \mathbf{1}\{\delta(P_i^{(m)}) \ge \frac{\delta}{2}\}] \ge 1 - 4|V|e^{-m(\frac{\delta}{2D})^2} - 4|E|e^{-m\gamma^2}.$$

Therefore, for a given $\eta > 0$ for $m \ge m_0$, where n_0 is the smallest solution of

$$\eta \ge 4|V|e^{-m(\frac{\delta}{2D})^2} + 4|E|e^{-m\gamma^2},$$

with probability at least $1-\eta$, we have

$$|\rho_m(G) - \rho_G| \le \gamma|E|D^{3/2}\frac{24}{\delta}.$$

We let $|\rho_m(G) - \rho_G| \le \epsilon$ and find γ as

$$\gamma = \frac{\epsilon}{|E|}\frac{\delta}{24}D^{-3/2}.$$

We plug this into the equation for m_0 to find

$$\eta \ge 4|V|e^{-m(\frac{\delta}{2D})^2} + 4|E|e^{-m\left(\frac{\epsilon}{|E|}\frac{\delta}{24}D^{-3/2}\right)^2}.$$

This leads to

$$\begin{aligned}m_0 &\ge \left(\frac{1}{\min\{(\frac{\delta}{2D})^2, (\frac{\epsilon}{|E|}\frac{\delta}{24}D^{-3/2})^2\}}\right)\log\left(\frac{8\max\{|V|,|E|\}}{\eta}\right) \\ &\ge \left(\frac{24|E|^2D^3}{\epsilon^2\delta^2}\right)\log\left(\frac{8\max\{|V|,|E|\}}{\eta}\right).\end{aligned}$$

10.10 Proof of Theorem 6

To prove Theorem 6, first we prove the following Lemma:

Lemma 4 *NMC optimization (4.1) can be written as follows:*

$$\begin{aligned} \max_{(i,i')\in E} \quad & \sum_{(i,i')\in E} \mathbb{E}[(\phi_i - \bar{\phi}_i)(\phi_{i'} - \bar{\phi}_{i'})] \\ & var\ (\phi_i) = 1, \quad 1 \le i \le n, \end{aligned} \tag{10.22}$$

where $\bar{\phi}_i$ and var (ϕ_i) represent the mean and the variance of the random variable ϕ_i.

Proof Denote the optimum value of optimization (10.22) by $\tilde{\rho}_G$. Let ϕ_i^* be an optimal solution of (4.1). The set of functions ϕ_i^* for $i = 1, \ldots, n$ is feasible for optimization (10.22) and therefore we have $\rho_G \le \tilde{\rho}_G$. On the other hand, let ϕ_i^{**} be an optimal solution of optimization (10.22). Let $\tilde{\phi}_i = \phi_i^{**} - \bar{\phi}_i^{**}$. The set of functions $\tilde{\phi}_i$ for $i = 1, \ldots, n$ is feasible for optimization (4.1). Thus, we have $\rho_G \ge \tilde{\rho}_G$. Therefore, we have that $\rho_G = \tilde{\rho}_G$. ■

Let $\mathbf{a}_i = \phi_i / \sqrt{\mathbf{p}_i}$. We have

$$1 = \text{var}(\phi_i) = \mathbb{E}[\phi_i^2] - (\mathbb{E}[\phi_i])^2 = \|\mathbf{a}_i\|_2^2 - (\mathbf{a}_i \sqrt{\mathbf{p}_i})^2 = \mathbf{a}_i^T \left(I - \sqrt{\mathbf{p}_i} \sqrt{\mathbf{p}_i}^T \right) \mathbf{a}_i.$$

We next show that the matrix $I - \sqrt{\mathbf{p}_i} \sqrt{\mathbf{p}_i}^T$ is positive semidefinite and the only vectors in its null space is $\mathbf{0}$ and $\sqrt{\mathbf{p}_i}$. This is because

$$\mathbf{x}^T \left(I - \sqrt{\mathbf{p}_i} \sqrt{\mathbf{p}_i}^T \right) \mathbf{x} = \|\mathbf{x}\|_2^2 - (\mathbf{x} \sqrt{\mathbf{p}_i})^2 \ge 0, \tag{10.23}$$

where we use Cauchy-Schwartz and $\|\sqrt{\mathbf{p}_i}\|_2^2 = 1$ to obtain the last inequality (10.23). This inequality becomes an equality if and only if $\mathbf{x} = 0$ or $\mathbf{x} = \sqrt{\mathbf{p}_i}$.
Now consider the objective function of optimization (10.22). We have

$$\begin{aligned} & \mathbb{E}[(\phi_i - \bar{\phi}_i)(\phi_{i'} - \bar{\phi}_{i'})] = \mathbb{E}[\phi_i \phi_{i'}] - \bar{\phi}_i \bar{\phi}_{i'} \\ & = \mathbf{a}_i^T Q_{ii'} \mathbf{a}_i - (\mathbf{a}_i^T \sqrt{\mathbf{p}_i})(\mathbf{a}_{i'}^T \sqrt{\mathbf{p}_{i'}}) = \mathbf{a}_i^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i} \sqrt{\mathbf{p}_{i'}}^T \right) \mathbf{a}_{i'}. \end{aligned}$$

Therefore, optimization (10.22) (which is equivalent to the NMC optimization (4.1) according to Lemma 4) can be written as,

$$\begin{aligned} \max_{\mathbf{a}_i} \quad & \sum_{(i,i')\in E} \mathbf{a}_i^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i} \sqrt{\mathbf{p}_{i'}}^T \right) \mathbf{a}_{i'} \\ & \mathbf{a}_i^T \left(I - \sqrt{\mathbf{p}_i} \sqrt{\mathbf{p}_i}^T \right) \mathbf{a}_i = 1, \quad 1 \le i \le n. \end{aligned} \tag{10.24}$$

For each i, since $I - \sqrt{\mathbf{p}_i} \sqrt{\mathbf{p}_i}^T$ is positive semidefinte. Thus, we can write $I - \sqrt{\mathbf{p}_i} \sqrt{\mathbf{p}_i}^T = B_i B_i^T$. Let $\mathbf{b}_i = B_i \mathbf{a}_i$. Thus, constraints of optimization (10.24) can be written as $\mathbf{b}_i^T \mathbf{b}_i = \|\mathbf{b}_i\|_2^2 = 1$.
We next write $\mathbf{a}_i$ as a function of $\mathbf{b}_i$. Note that since B_i is not invertible, there are many choices for $\mathbf{a}_i$ as a function of $\mathbf{b}_i$ characterized as follows: Let $U_i \Sigma_i U_i^T$ be the singular value decomposition of B_i. The vector $\sqrt{\mathbf{p}_i}$ is the singular vector corresponding to singular value zero ($\sigma_i^{(1)} = 0$).

$$\mathbf{a}_i = \left([U_i^{(2)}, \ldots, U_i^{(|\mathcal{X}_i|)}] \text{diag}(1/\sigma_i^2, \ldots, 1/\sigma_i^{(n_i)}) [U_i^{(2)}, \ldots, U_i^{(|\mathcal{X}_i|)}]^T \right) \mathbf{b}_i + \alpha_i \sqrt{\mathbf{p}_i} = A_i \mathbf{b}_i + \alpha_i \sqrt{\mathbf{p}_i}, \tag{10.25}$$

where $U_i^{(j)}$ is the j-th column of U_i, $\sigma_i^{(j)}$ is the j-th singular value of B_i, and α_i can be any scalar.[3] Below, we show that all choices of $\mathbf{a}_i$ according to (10.25) lead to the same objective function of optimization (10.24):

$$\begin{aligned}
\mathbf{a}_i^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_{i'}}^T\right)\mathbf{a}_{i'} &= \mathbf{b}_i^T A_i^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_{i'}}^T\right) A_{i'}\mathbf{b}_{i'} \\
&+ \mathbf{b}_i^T A_i^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_{i'}}^T\right) \alpha_{i'}\sqrt{\mathbf{p}_{i'}} \\
&+ \alpha_i \sqrt{\mathbf{p}_i}^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_{i'}}^T\right) A_{i'}\mathbf{b}_{i'} \\
&+ \alpha_i \sqrt{\mathbf{p}_i}^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_{i'}}^T\right) \alpha_{i'}\sqrt{\mathbf{p}_{i'}} \\
&= \mathbf{b}_i^T A_i^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_{i'}}^T\right) A_{i'}\mathbf{b}_{i'} \\
&+ \mathbf{b}_i^T A_i^T Q_{ii'} \alpha_{i'}\sqrt{\mathbf{p}_{i'}} \\
&+ \alpha_i \sqrt{\mathbf{p}_i}^T Q_{ii'} A_{i'}\mathbf{b}_{i'} \\
&+ \alpha_i \alpha_{i'}(1-1) \\
&= \mathbf{b}_i^T A_i^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_{i'}}^T\right) A_{i'}\mathbf{b}_{i'} \\
&+ \mathbf{b}_i^T A_i^T \sqrt{\mathbf{p}_i}\alpha_{i'} \\
&+ \alpha_i \sqrt{\mathbf{p}_{i'}}^T A_{i'}\mathbf{b}_{i'} = \mathbf{b}_i^T A_i^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_{i'}}^T\right) A_{i'}\mathbf{b}_{i'}.
\end{aligned}$$

Therefore, the NMC optimization (4.1) can be written as

$$\begin{aligned}
\max_{\mathbf{b}_i} \quad & \sum_{(i,i')\in E} \mathbf{b}_i^T A_i^T \left(Q_{ii'} - \sqrt{\mathbf{p}_i}\sqrt{\mathbf{p}_{i'}}^T\right) A_{i'}\mathbf{b}_{i'} \\
& \|\mathbf{b}_i\|_2 = 1.
\end{aligned}$$

This completes the proof.

10.11 Proof of Proposition 6

We use the following lemma in the proof of Proposition 6.

Lemma 5 *Let X and Y be two random variables such that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. The solution to the optimization*

$$\begin{aligned}
\max_{\phi} \quad & \mathbb{E}[\phi(X)Y] \\
& \mathbb{E}[\phi(X)] = 0 \\
& \mathbb{E}[\phi(X)^2] = 1,
\end{aligned}$$

is $\phi(X) = \frac{\mathbb{E}[Y|X]}{\sqrt{\mathbb{E}[(\mathbb{E}[Y|X])^2]}}$.

[3] Since B_i is symmetric it has a set of $|\mathcal{X}_i|$ orthonormal eigenvectors and can be written as

$$B_i = \sum_{j=1}^{|\mathcal{X}_i|} \mathbf{v}_j \sigma_i^{(j)} \mathbf{v}_j^T.$$

Let $\mathbf{b}_i = \sum_{j=2}^{|\mathcal{X}_i|} \beta_j \mathbf{v}_j$ and $\mathbf{a}_i = \sum_{j=1}^{|\mathcal{X}_i|} \alpha_j \mathbf{v}_j$. From $\mathbf{b}_i = B_i \mathbf{a}_i$, we obtain that $\alpha_j = \beta_j / \sigma_i^{(j)}$ for $j \geq 2$, where α_1 can be any scalar.

Proof We have

$$\mathbb{E}[\phi(X)Y] = \mathbb{E}\left[\phi(X)\mathbb{E}[Y|X]\right] \leq \frac{1}{2}\left(\mathbb{E}[\mathbb{E}[Y|X]^2] + \mathbb{E}[\phi(X)^2]\right),$$

where we use $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y] = 0$ to derive the last equality. Note that the inequality becomes an equality if and only if $\phi(X) = \mathbb{E}[Y|X]$, which completes the proof. ■

At each iteration of Algorithm 3, we fix all the functions except one of them and then find the optimum of that function. We show that the objective of optimization (4.1) increases at each step of Algorithm (3). Since the objective is bounded, the convergence follows. Without loss of generality, we show that by fixing all functions for random variables $X_2, \ldots, X_n$ and updating the corresponding function of random variable X_1, the objective of optimization (4.1) increases (or does not change). We have

$$\begin{aligned}
&\text{argmax}_{\phi_1(X_1)} \sum_{(i,j)\in E} \mathbb{E}[\phi_i(X_i)\phi_j(X_j)] \\
=&\text{argmax}_{\phi_1(X_1)} \mathbb{E}\left[\left(\sum_{j\in\mathcal{N}_i} \phi_j(X_j)\right)\phi_i(X_i)\right],
\end{aligned}$$

where $\mathbb{E}[\phi_1] = 1$ and $\mathbb{E}[\phi_1^2] = 1$. Using Lemma 5, the update of $\phi_1(X)$ is

$$\phi_1(X_1) = \frac{\mathbb{E}\left[\left(\sum_{j\in\mathcal{N}(i)} \phi_j(X_j)\right)|X_1\right]}{\mathbb{E}\left[\left(\mathbb{E}\left[\left(\sum_{j\in\mathcal{N}(i)} \phi_j(X_j)\right)|X_1\right]\right)^2\right]}.$$

This concludes the proof.

10.12 Proof of Theorem 8

For any realization of the partitioning, consider NMC over all sub-graphs G_m $(1 \leq m \leq M)$ and denote the corresponding functions by $\hat{\phi}_i$ for $1 \leq i \leq n$. We have

$$\begin{aligned}
\rho_G &= \sum_{(i,j)\in E} \mathbb{E}[\phi_i^*\phi_j^*] = \sum_{(i,j)\in E\setminus E^c} \mathbb{E}[\phi_i^*\phi_j^*] + \sum_{(i,j)\in E^c} \mathbb{E}[\phi_i^*\phi_j^*] \\
&= \sum_{m=1}^{M} \sum_{(i,j)\in E_m} \mathbb{E}[\phi_i^*\phi_j^*] + \sum_{(i,j)\in E^c} \mathbb{E}[\phi_i^*\phi_j^*] \\
&\leq \sum_{m=1}^{M} \hat{\rho}_{G_m} + \sum_{(i,j)\in E^c} \mathbb{E}[\phi_i^*\phi_j^*] \\
&= \hat{\rho}_G + \sum_{(i,j)\in E} \mathbf{1}\{(i,j) \in E^c\}\mathbb{E}[\phi_i^*\phi_j^*].
\end{aligned}$$

Therefore, by taking expectation over the partitioning, we obtain

$$\rho_G \leq \mathbb{E}[\hat{\rho}_G] + \epsilon\rho_G,$$

which gives us

$$(1-\epsilon)\rho_G \leq \mathbb{E}[\hat{\rho}_G].$$

This completes the proof.

10.13 Proof of Theorem 9

Consider the following optimization:

$$\max_{\mathbf{a}_i} \quad \sum_{(i,i')\in E} \sum_{j=2}^{K} a_{i,j} a_{i',j} \ \rho_{i,i'}^{j} \qquad (10.26)$$
$$\sum_{j=2}^{\infty} (a_{i,j})^2 = 1, \quad 1 \le i \le n.$$

where we only consider $2 \le j \le K$. We prove Theorem 9 for any $K \ge 2$. Thus, since $\rho_{i,i'}^{K} \to 0$ as $k \to \infty$ for all $i \ne i'$, Theorem 9 holds. Let Λ be the matrix of correlation coefficients where $\Lambda[i,i'] = \rho_{i,i'}$. Diagonal elements of Λ are all zero, as we ignore self-loops. Define

$$\mathbf{x} = [a_{1,1}, a_{2,1}, \ldots, a_{n,1}, a_{1,2}, \ldots, a_{n,K}]^T. \qquad (10.27)$$

Moreover, define A_0 as an $nK \times nK$ matrix composed of K^2 blocks of size $n \times n$ where its m-th diagonal block is equal to $-2\ \Lambda^{\cdot m}$, where $A^{\cdot m}[i,j] \triangleq A[i,j]^m$. Off-diagonal blocks of A_0 are all zeros. Moreover, define A_i for $1 \le i \le n$ as an $nK \times nK$ matrix where $A_i[m\ i, m\ i] = 1$ for $1 \le m \le K$, otherwise it is zero. Therefore, optimization (10.26) can be re-written as the following standard quadratic optimization:

$$\min_{\mathbf{x}} \quad \frac{1}{2}\mathbf{x}^T A_0 \mathbf{x} \qquad (10.28)$$
$$\frac{1}{2}\mathbf{x}^T\ A_i\ \mathbf{x} - \frac{1}{2} \le 0, \quad 1 \le i \le n.$$

Note that equality constraints of optimization (10.26) are replaced by inequality ones in optimization (10.28). This is because, since $\rho_{i,i'}^2 \ge 0$, optimal solutions of optimization (10.28) occur in the boundary of its feasible set. Optimization (10.28) is a non-convex quadratic minimization problem with quadratic constraints. Reference [63] characterizes necessary and sufficient conditions for global minimizers of a generalized form of optimization (10.28). Let

$$\bar{\mathbf{x}} = [s_1, s_2, \ldots, s_n, 0, \ldots, 0]^T, \qquad (10.29)$$

where $s_i \in \{-1, 1\}$, for $1 \le i \le n$. According to reference [63], to have $\bar{\mathbf{x}}$ as a global minimizer of optimization (10.28), we need to have

$$\Big(\sum_{i=1}^{n} \lambda_i A_i + A_0\Big)\bar{\mathbf{x}} = 0 \qquad (10.30)$$

and

$$\sum_{i=1}^{n} \lambda_i A_i + A_0 \succeq 0, \qquad (10.31)$$

where $\lambda_i \ge 0$, and $A \succeq 0$ means that A is a positive semi-definite matrix. Using definitions of A_0, A_i, and $\bar{\mathbf{x}}$, equation (10.30) is satisfied iff

$$\lambda_i = 2 \sum_{i' \ne i} s_i s_{i'} \rho_{i,i'} \ge 0, \quad 1 \le i \le n. \qquad (10.32)$$

Using (10.32) and Gerschgorin's circle theorem, if

$$\sum_{i' \neq i} (1 - s_i s_{i'}) \rho_{i,i'} \geq 0, \quad \forall 1 \leq i \leq n, \tag{10.33}$$

$$\sum_{i' \neq i} s_i s_{i'} \rho_{i,i'} \geq \sum_{i' \neq i} \rho_{i,i'}^2, \quad \forall 1 \leq i \leq n, \tag{10.34}$$

conditions (10.31) are satisfied. Thus, $\bar{\mathbf{x}}$ is a global minimizer of optimization (10.28). This completes the proof.

10.14 Proof of Proposition 9

Under assumptions of Theorem 9 and using the definition of Hermitte-Chebychev polynomials (2.8), we can restrict the feasible set of optimization (3) to the set of functions $\phi_i(X_i) = s_i X_i$ where $s_i \in \{-1, 1\}$ for all $1 \leq i \leq n$. Moreover, we have

$$\mathbb{E}[\phi_i(X_i)\phi_{i'}(X_{i'})] = s_i s_{i'} \rho_{i,i'}.$$

Furthermore, $\mathbb{E}[s_i X_i] = \mathbb{E}[X_i] = 0$, and $\mathbb{E}\big[(s_i X_i)^2\big] = \mathbb{E}[X_i^2] = 1$. This completes the proof.

10.15 Proof of Theorem 11

We re-write optimization (6.7) as follows:

$$\begin{aligned} \max \sum_{(i,i')} & \mathbb{E}[g_i(f_i(X_i))\ g_{i'}(f_{i'}(X_{i'}))], \\ & \mathbb{E}[g_i(f_i(X_i))] = 0, \quad 1 \leq i \leq n, \\ & \mathbb{E}[g_i(f_i(X_i)^2] = 1, \quad 1 \leq i \leq n. \end{aligned} \tag{10.35}$$

Define $\phi_i(X_i) = g_i(f_i(X_i))$ for $1 \leq i \leq n$. Since f_i's are bijective and differentiable, feasible regions of optimizations (10.35) and (3) are equal. Under the assumptions of Corrolary 1, $\phi_i^*(X_i) = X_i$. Thus, $g_i^*(f_i(X_i)) = X_i$. Thus, according to Theorem 10, if $(i, j) \notin E_{nmc}$, then,

$$X_i \perp X_j | \{X_k, k \neq i, j\}. \tag{10.36}$$

Since Y_i's are bijective functions of X_i's, this completes the proof.

11 Acknowledgements

We thank Yue Li and Gerald Quon for discussions regarding the cancer application. We also thank Yue Li for providing the cancer datasets. We thank Flavio du Pin Calmon and Vahid Montazerhojjat for helpful discussions.

References

[1] R. Albert, "Network inference, analysis, and modeling in systems biology," *The Plant Cell*, vol. 19, no. 11, pp. 3327–3338, 2007.

[2] M. S. Handcock, *Statistical models for social networks: Inference and degeneracy.* National Academies Press, 2002.

[3] A. G. Haldane, "Rethinking the financial network," *Speech delivered at the Financial Student Association, Amsterdam, April*, pp. 1–26, 2009.

[4] S. H. Strogatz, *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering.* Westview press, 2014.

[5] T. M. Cover and J. A. Thomas, *Elements of information theory.* John Wiley & Sons, 2012.

[6] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[7] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[8] H. Gebelein, "Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *ZAMM-Journal of Applied Mathematics and Mechanics*, vol. 21, no. 6, pp. 364–379, 1941.

[9] H. O. Hirschfeld, "A connection between correlation and contingency," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 31, no. 04, 1935, pp. 520–524.

[10] O. Sarmanov, "Maximum correlation coefficient (nonsymmetric case)," *Selected Translations in Mathematical Statistics and Probability*, vol. 2, pp. 207–210, 1962.

[11] A. Rényi, "On measures of dependence," *Acta mathematica hungarica*, vol. 10, no. 3, pp. 441–451, 1959.

[12] M. J. Greenacre, *Theory and applications of correspondence analysis.* Academic Press, 1984.

[13] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 580–598, 1985.

[14] H. O. Lancaster, "Some properties of the bivariate normal distribution considered in the form of a contingency table," *Biometrika*, vol. 44, no. 1-2, pp. 289–292, 1957.

[15] Y. Polyanskiy, "Hypothesis testing via a comparator and hypercontractivity," *Prob. Peredachi Inform*, 2013.

[16] T. Courtade, "Outer bounds for multiterminal source coding via a strong data processing inequality," in *IEEE International Symposium on Information Theory*, 2013, pp. 559–563.

[17] F. Calmon, M. Varia, and M. Médard, "An exploration of the role of principal inertia components in information theory," in *Information Theory Workshop (ITW), 2014 IEEE*, 2014, pp. 252–256.

[18] F. Calmon, M. Médard, M. Varia, K. R. Duffy, M. M. Christiansen, and L. M. Zeger, "Hiding symbols and functions: New metrics and constructions for information-theoretic security," *arXiv preprint:1503.08513*, 2015.

[19] F. Calmon, M. Varia, and M. Medard, "On information-theoretic metrics for symmetric-key encryption and privacy," in *Allerton Conference on Communication, Control, and Computing*, 2014, pp. 889–894.

[20] F. Calmon, M. Varia, M. Médard, M. M. Christiansen, K. R. Duffy, and S. Tessaro, "Bounds on inference," in *Allerton Conference on Communication, Control, and Computing*, 2013, pp. 567–574.

[21] M. Raginsky, "Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels," *arXiv preprint:1411.3575*, 2014.

[22] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover," *arXiv preprint:1304.6133*, 2013.

[23] M. T. Chu and J. L. Watterson, "On a multivariate eigenvalue problem, part i: Algebraic theory and a power method," *SIAM Journal on Scientific Computing*, vol. 14, no. 5, pp. 1089–1106, 1993.

[24] H. Hotelling, "The most predictable criterion." *Journal of Educational Psychology*, vol. 26, no. 2, p. 139, 1935.

[25] ——, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.

[26] L.-H. Zhang and M. T. Chu, "Computing absolute maximum correlation," *IMA Journal of Numerical Analysis*, vol. 32, no. 1, pp. 163–184, 2012.

[27] M. R. Garey, D. S. Johnson, and L. Stockmeyer, "Some simplified NP-complete graph problems," *Theoretical computer science*, vol. 1, no. 3, pp. 237–267, 1976.

[28] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *Journal of the ACM (JACM)*, vol. 42, no. 6, pp. 1115–1145, 1995.

[29] M. Newman, *Networks: an introduction*. Oxford University Press, 2010.

[30] Y. Li and Z. Zhang, "Potential microrna-mediated oncogenic intercellular communication revealed by pan-cancer analysis," *Scientific reports*, vol. 4, 2014.

[31] H. S. Witsenhausen, "On sequences of pairs of dependent random variables," *SIAM Journal on Applied Mathematics*, vol. 28, no. 1, pp. 100–113, 1975.

[32] W. Bryc and A. Dembo, "On the maximum correlation coefficient," *Theory of Probability and its Applications*, vol. 49, no. 1, pp. 132–138, 2005.

[33] E. M. Stein and R. Shakarchi, *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.

[34] D. P. Bertsekas, *Nonlinear programming*. Athena scientific, 1999.

[35] P. Horst, *Factor analysis of data matrices*. Holt, Rinehart and Winston Publisher, 1965.

[36] ——, "Relations among sets of measures," *Psychometrika*, vol. 26, no. 2, pp. 129–149, 1961.

[37] L.-H. Zhang and L.-Z. Liao, "An alternating variable method for the maximal correlation problem," *Journal of Global Optimization*, vol. 54, no. 1, pp. 199–218, 2012.

[38] G. H. Golub and C. F. Van Loan, *Matrix Computations*. JHU Press, 2012, vol. 3.

[39] K. Jung and D. Shah, "Local algorithms for approximate inference in minor-excluded graphs," in *Advances in Neural Information Processing Systems*, 2008, pp. 729–736.

[40] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[41] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied linear statistical models*. Irwin Chicago, 1996, vol. 4.

[42] K. Baba, R. Shibata, and M. Sibuya, "Partial correlation and conditional correlation as measures of conditional independence," *Australian and New Zealand Journal of Statistics*, vol. 46, no. 4, pp. 657–664, 2004.

[43] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *The Journal of Machine Learning Research*, vol. 9, pp. 485–516, 2008.

[44] B. W. Stewart, P. Kleihues, and et al., *World cancer report*. IARC press Lyon, 2003, vol. 57.

[45] G. A. Calin and C. M. Croce, "MicroRNA signatures in human cancers," *Nature Reviews Cancer*, vol. 6, no. 11, pp. 857–866, 2006.

[46] W. Huber, A. Von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression," *Bioinformatics*, vol. 18, no. suppl 1, pp. S96–S104, 2002.

[47] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[48] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[49] J. M. Bland and D. G. Altman, "The logrank test," *BMJ*, vol. 328, no. 7447, p. 1073, 2004.

[50] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[51] G. S. Bailey, A. P. Reddy, C. B. Pereira, U. Harttig, W. Baird, J. M. Spitsbergen, J. D. Hendricks, G. A. Orner, D. E. Williams, and J. A. Swenberg, "Nonlinear cancer response at ultralow dose: a 40800-animal ed001 tumor and biomarker study," *Chemical research in toxicology*, vol. 22, no. 7, pp. 1264–1276, 2009.

[52] X. Huang, W. Qian, I. H. El-Sayed, and M. A. El-Sayed, "The potential use of the enhanced nonlinear properties of gold nanospheres in photothermal cancer therapy," *Lasers in surgery and medicine*, vol. 39, no. 9, pp. 747–753, 2007.

[53] J. S. Lowengrub, H. B. Frieboes, F. Jin, Y. Chuang, X. Li, P. Macklin, S. Wise, and V. Cristini, "Nonlinear modelling of cancer: bridging the gap between cells and tumours," *Nonlinearity*, vol. 23, no. 1, p. R1, 2010.

[54] L. Nagel and R. Rohrer, "Computer analysis of nonlinear circuits, excluding radiation (cancer)," *IEEE Journal of Solid-State Circuits*, vol. 6, no. 4, pp. 166–182, 1971.

[55] J. G. Wagner, J. W. Gyves, P. L. Stetson, S. C. Walker-Andrews, I. S. Wollner, M. K. Cochran, and W. D. Ensminger, "Steady-state nonlinear pharmacokinetics of 5-fluorouracil during hepatic arterial and intravenous infusions in cancer patients," *Cancer research*, vol. 46, no. 3, pp. 1499–1506, 1986.

[56] D. M. Bates and D. G. Watts, *Nonlinear regression: iterative estimation and linear approximations*. Wiley Online Library, 1988.

[57] J. Hasty, D. McMillen, F. Isaacs, and J. J. Collins, "Computational studies of gene regulatory networks: in numero molecular biology," *Nature Reviews Genetics*, vol. 2, no. 4, pp. 268–279, 2001.

[58] J. Oncley, E. Ellenbogen, D. Gitlin, and F. Gurd, "Protein–protein interactions," *The Journal of Physical Chemistry*, vol. 56, no. 1, pp. 85–92, 1952.

[59] B. Golub and M. O. Jackson, "Naive learning in social networks and the wisdom of crowds," *American Economic Journal: Microeconomics*, pp. 112–149, 2010.

[60] L. Mirsky, "Symmetric gauge functions and unitarily invariant norms," *The quarterly journal of mathematics*, vol. 11, no. 1, pp. 50–59, 1960.

[61] L. Devroye, "The equivalence of weak, strong and complete convergence in l_1 for kernel density estimates," *The Annals of Statistics*, pp. 896–904, 1983.

[62] D. Berend and A. Kontorovich, "On the convergence of the empirical distribution," *arXiv preprint:1205.6711*, 2012.

[63] V. Jeyakumar, A. M. Rubinov, and Z. Wu, "Non-convex quadratic minimization problems with quadratic constraints: global optimality conditions," *Mathematical Programming*, vol. 110, no. 3, pp. 521–541, 2007.