ROBUST IDENTIFICATION OF GRAPH STRUCTURE

Georgios V. Karanikolas and Georgios B. Giannakis

Dept. of ECE, University of Minnesota, Minneapolis, MN

ABSTRACT

Partial correlations (PCs) and the related inverse covariance matrix adopted by graphical lasso, are widely applicable tools for learning graph connectivity given nodal observations. The resultant estimators however, can be sensitive to outliers. Robust approaches developed so far to cope with outliers do not (explicitly) account for nonlinear interactions possibly present among nodal processes. This can hurt the identification of graph connectivity, merely due to model mismatch. To overcome this limitation, a novel formulation of robust PC is introduced based on nonlinear kernel functions. The proposed scheme leverages robust ridge regression techniques, spectral Fourier feature based kernel approximants, and robust association measures. Numerical tests on synthetic and real data illustrate the potential of the novel approach.

Index Terms— Robust statistics, kernel-based methods, network topology inference

1. INTRODUCTION

Inferring the structure of a graph from nodal observations is an important task in a wide gamut of disciplines including genomics, where regulation relations among genes are learned from gene expression profiles, and neuroscience, where relations among regions of the brain are deduced from time series acquired by various brain imaging modalities [5, 15, 2].

PC-based structure estimators aim at introducing an edge between a pair of nodes only if their relation is a direct one, that is not one mediated through additional nodes. This is accomplished by partialing out the effect of the remaining nodes by means of linear regression [11]. In the absence of nonlinear mediation effects and outliers, (linear) PC achieves its goal. Nonlinear influences are widely present though, and they can be found, for instance, in the context of interaction modeling among species in ecological networks [4]. Real world data is also oftentimes contaminated with outlying observations. Finally, note that the presence of nonlinearities and outliers are, of course, by no means mutually exclusive.

Prior works. Robust approaches to learning PC-based connectivity either directly estimate the PC coefficients between pairs of nodes [19], or estimate the inverse covariance matrix instead[1] via robust variants of the graphical lasso (GL) [24, 7, 3]. Both classes of methods however, are not designed to account for nonlinear mediation effects. The former explicitly rely on linear regression. The latter typically boil down to using a robust estimate of some first-order central product-moment between nodal observations (such as the covariance or correlation) as the input, in place of the covariance matrix, to a GL problem; see e.g., [3, 24, 7]. Using only first-order information however, is generally not sufficient for modeling nonlinear mediating dependencies among nodal observations. On the other hand, the available nonlinear PC approaches are not designed to be robust to outliers [10].

To overcome the limitations of existing approaches, the present work introduces a *robust* PC criterion that also accounts for *nonlinear* mediating influences among nodes. To that end, first, nonlinear modeling functions are employed, instead of the linear ones considered by plain PC. Nonparametric nonlinear regression estimators are then approximated by parametric ones, through the use of spectral Fourier features, what enables scalability. To limit the effect of 'outliers,' a highly robust ridge regression scheme will be leveraged. As the goal of this scheme is to obtain an estimator that accurately describes 'inliers,' outlying fit residuals may be obtained for outlying observations. A robust counterpart of the Pearson correlation coefficient shall thus be leveraged in order to correctly assess the association between said residuals, and finally obtain our robust partial correlation measure.

Contributions. The present work is the first to jointly address the presence of *nonlinear* mediation effects and *outliers* in the context of PC-based learning of graph structure. As part of the novel approach, a highly robust and scalable approximate kernel ridge regression scheme is also outlined.

Notation. The all-ones vector is denoted by $\mathbf{1}$, while $\mathbf{I}$ stands for the identity matrix. The indicator function is given by $\mathbb{1}\{\cdot\}$, and $\lfloor a \rfloor$ denotes the largest integer that is less than or equal to a. The binomial coefficient is given by $\binom{n}{k}$, and $[T]$ is shorthand for the set $\{1, 2, \ldots, T\}$. Finally, $\|\mathbf{A}\|_F$ stands for the Frobenius norm of matrix $\mathbf{A}$, $\|\mathbf{x}\|_p$ denotes the ℓ_p norm of the vector $\mathbf{x}$, and $\mathbf{x}^\top$ stands for its transpose.

This work was supported by NSF grants 1901134, 2126052, 2128593, 2212318, 2220292, 2312547.

[1] Linear PCs are expressible in terms of inverse covariance matrix entries.

2. PRELIMINARIES

Consider a network of multiple interconnected agents with each agent represented by a vertex (node) $\nu \in \mathcal{V}$, where $\mathcal{V}$ denotes the set of all nodes. Node ν is represented by a vector of observations (features) $\mathbf{x}_\nu := [x_\nu[1], \ldots, x_\nu[T]]^\top$. For each pair of nodes (i, j), we wish to assess the association between the corresponding nodal vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, while accounting also for the (i, j) connectivity through the remaining nodes [11]. Our ultimate goal is outlier-resilient identification of these pairwise associations, as they collectively yield a robust (weighted) estimate of the overall graph connectivity.

To define the (sample) partial correlation coefficient, let $\hat{\mathbf{x}}_{i|\mathcal{V}\setminus ij}, \hat{\mathbf{x}}_{j|\mathcal{V}\setminus ij}$ denote the estimated vectors at nodes $i, j \in \mathcal{V}$ using linear regression of all but the (i, j) nodes [11]. In particular, the regression model postulated in PC for node i is $x_i[t] = f_{i|\mathcal{V}\setminus ij}(\boldsymbol{\chi}_{\setminus ij}[t]) + \epsilon_{i|\mathcal{V}\setminus ij}[t]$, where $f_{i|\mathcal{V}\setminus ij}$ is a linear function and $\boldsymbol{\chi}_{\setminus ij}[t]$ collects the observations at all nodes except for i, j at time t. As we will be focusing on the arbitrary pair (i, j), we will selectively drop $_{|\mathcal{V}\setminus ij}$ from our notation hereafter. Letting $\tilde{\mathbf{x}}_i := \mathbf{x}_i - \hat{\mathbf{x}}_i$ and similarly for $\tilde{\mathbf{x}}_j$, the sample PC coefficient of $\mathbf{x}_i, \mathbf{x}_j$ with respect to the observations at the remaining nodes $\{\mathbf{x}_k\}_{k\in\mathcal{V}/ij}$ is given by

$$\hat{\varrho}_{ij} := \frac{(\tilde{\mathbf{x}}_i - \bar{\tilde{\mathbf{x}}}_i)^\top (\tilde{\mathbf{x}}_j - \bar{\tilde{\mathbf{x}}}_j)}{\|\tilde{\mathbf{x}}_i - \bar{\tilde{\mathbf{x}}}_i\|_2 \, \|\tilde{\mathbf{x}}_j - \bar{\tilde{\mathbf{x}}}_j\|_2} \tag{1}$$

where $\bar{\tilde{\mathbf{x}}}_i := T^{-1} \sum_{t=1}^T \tilde{x}_i[t]\mathbf{1}$. Note that (1) corresponds to the Pearson correlation coefficient between the residual vectors $\{\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j\}$. The magnitude $|\hat{\varrho}_{ij}|$ indicates the strength of the association purely between nodes i and j, and not through the remaining nodes. Deciding whether the (i, j) edge is present amounts to performing a hypothesis test with $\hat{\varrho}_{ij}$ (or a function thereof) as the test statistic.

3. PROPOSED APPROACH

The linear nature of f_i makes it ill-suited for settings where nonlinearities are present. In the spirit of [10], that however does not deal with outlier-robust estimators, a nonlinear f_i, estimated by means of kernel-based regression, will be leveraged in order to account for *nonlinear* mediating interactions. Moreover, a spectral Fourier feature based approximation for the kernel will be used here, as this effects graceful scaling of the complexity with respect to T in the ensuing estimation tasks [18]. This is important, as it is the norm for robust regression estimators to be obtained via iterative algorithms that perform multiple passes over the data [9].

In short, consider a 'stationary' kernel κ that satisfies $\kappa(\boldsymbol{\chi}, \boldsymbol{\chi}') = \kappa(\boldsymbol{\chi} - \boldsymbol{\chi}')$, and let $\bar{\kappa}(\breve{\boldsymbol{\chi}}) = \bar{\kappa}(\boldsymbol{\chi} - \boldsymbol{\chi}')$ denote its normalized version. Drawing i.i.d. random vectors $\{\boldsymbol{\omega}_i\}_{i=1}^m$ from the probability density function $\pi_\kappa(\boldsymbol{\omega}) = \mathcal{F}(\bar{\kappa}(\breve{\boldsymbol{\chi}}))$, where $\mathcal{F}$ denotes the Fourier transform, and letting $\phi(\boldsymbol{\chi}) = \frac{1}{\sqrt{m}}[\cos(\boldsymbol{\omega}_1^\top \boldsymbol{\chi}) \;\; \sin(\boldsymbol{\omega}_1^\top \boldsymbol{\chi}) \;\; \ldots \cos(\boldsymbol{\omega}_m^\top \boldsymbol{\chi}) \;\; \sin(\boldsymbol{\omega}_m^\top \boldsymbol{\chi})]^\top$ a kernel approximant is obtained as $\breve{\kappa}(\boldsymbol{\chi}, \boldsymbol{\chi}') = \phi^\top(\boldsymbol{\chi})\phi(\boldsymbol{\chi}')$, where $\phi(\boldsymbol{\chi})$ is the so-termed spectral feature vector of $\boldsymbol{\chi}$ [12]; see also [18] for approximation guarantees.

Vector $\phi(\boldsymbol{\chi})$ yields a linear parametric approximant of f_i given by $\breve{f}_i(\boldsymbol{\chi}_{/ij}[t]) = \boldsymbol{\theta}_i^\top \phi(\boldsymbol{\chi}_{/ij}[t])$ for the nonlinear function, where $\boldsymbol{\theta}_i \in \mathbb{R}^{2m}$ is the vector of regression coefficients [18]. Estimating $\breve{f}_i$ amounts to estimating $\boldsymbol{\theta}_i$ at reduced complexity $\mathcal{O}(Tm^2)$ compared to $\mathcal{O}(T^3)$ incurred by (exact) kernel ridge regression, with m chosen such that $m \ll T$.

Before introducing our robust spectral feature based estimator for f_i, it is important to outline robust (M-)estimators of scale and the associated (ρ-)functions. The M-estimator $\hat{\sigma}_\rho(\mathbf{r})$ for the observations $\mathbf{r} := [r_1, \ldots, r_T]^\top$ is given by the solution with respect to σ of

$$\frac{1}{T}\sum_{t=1}^T \rho\left(\frac{r_t}{\sigma}\right) = \delta \tag{2}$$

where $\rho(z)$ is a chosen ρ-function (nondecreasing in $|z|$, increasing for $z > 0$ with $\rho(z) < \rho(\infty) := \lim_{z\to\infty} \rho(z)$, and with $\rho(0) = 0$), and δ is a constant determining the breakdown point (BDP)[2] of $\hat{\sigma}_\rho(\mathbf{r})$ [14]. If one were to replace $\rho(z)$ with z^2 and let $\delta = 1$, the solution of (2) will boil down to the sample root mean square of $\mathbf{r}$, which is not robust.

In general, the choice of $\rho(\cdot)$ determines the effect outliers have on $\hat{\sigma}_\rho(\mathbf{r})$. In order to obtain a highly robust estimator, the effect of large outliers should be zero [14]. This is accomplished by using a *bounded* ρ-function, such as the bisquare $\rho_c(z) := 1 - (1 - (z/c)^2)^3 \, \mathbb{1}\{|z| \le c\}$, with c being a tuning constant. The bisquare is a standard choice in robust statistics [14, 26], and will be relied upon hereafter. As $\text{BDP}(\hat{\sigma}_\rho(\mathbf{r})) = \min(\delta/\rho(\infty), 1-\delta/\rho(\infty))$, it follows that the bisquare with $\delta = 0.5$ satisfies $\text{BDP}(\hat{\sigma}_{\rho_c}(\mathbf{r})) = 0.5$, whereas for unbounded ρ-functions, such as the Huber or ℓ_1 loss, the BDP reduces to zero [14].

With spectral features $\phi(\boldsymbol{\chi})$ as regressors, a non-robust regularized estimate of $\boldsymbol{\theta}_i$ is given by

$$\hat{\boldsymbol{\theta}}_i \;=\; \underset{\boldsymbol{\theta}_i\in\mathbb{R}^{2m}}{\arg\min} \sum_{t=1}^T (x_i[t] - \boldsymbol{\theta}_i^\top \phi(\boldsymbol{\chi}_{\setminus ij}[t]))^2 + \lambda\|\boldsymbol{\theta}_i\|_2^2 \tag{3}$$

with λ controlling the regularization strength. Let $r_{it}(\boldsymbol{\theta}_i) := x_i[t] - \boldsymbol{\theta}_i^\top \phi(\boldsymbol{\chi}_{\setminus ij}[t])$ be the residual for observation $x_i[t]$ corresponding to the (arbitrary) choice of regression coefficients $\boldsymbol{\theta}_i$, and define $\mathbf{r}_i(\boldsymbol{\theta}_i) := [r_{i1}(\boldsymbol{\theta}_i), \ldots, r_{iT}(\boldsymbol{\theta}_i)]^\top$.

A robust counterpart of (3) can be obtained as

$$\hat{\boldsymbol{\theta}}_i \;\in\; \underset{\boldsymbol{\theta}_i\in\mathbb{R}^{2m}}{\arg\min} \;\; \hat{\sigma}_{(i)}^2 \sum_{t=1}^T \rho_{c_1}\left(\frac{r_{it}(\boldsymbol{\theta}_i)}{\hat{\sigma}_{(i)}}\right) \;+\; \lambda\|\boldsymbol{\theta}_i\|_2^2 \tag{4}$$

where $\hat{\sigma}_{(i)}$ is an initial estimate of scale of the residuals, required for defining $\hat{\boldsymbol{\theta}}_i$ [13, 14]. In particular, letting $\hat{\boldsymbol{\theta}}_i^{(0)}$ be an

[2] At a high level, the BDP of an estimator is the smallest fraction of observations that when replaced by outliers can take the estimator outside of any bounded set; see [14] for formal definitions. Scale estimators are additionally required to remain bounded away from zero.

initial estimate of $\boldsymbol{\theta}_i$, and $\mathbf{r}_i(\hat{\boldsymbol{\theta}}_i^{(0)})$ be the associated residuals, $\hat{\sigma}_{(i)}$ is defined as the M-estimator of scale for said residuals, that is $\hat{\sigma}_{(i)} := \hat{\sigma}_{\rho_{c_0}}(\mathbf{r}_i(\hat{\boldsymbol{\theta}}_i^{(0)}))$ [13]. Notice that upon replacing $\rho_{c_1}(z)$ with z^2 in (4), one recovers (3). Finally, the tuning constant c_1 is chosen so as to achieve an efficiency of 0.85 at normal data; see [14] for details.

Remark 1. Thanks to its boundedness, it turns out that the bisquare is highly robust to outliers even as $\lambda \to 0$ to minimize bias [17]. In contrast, unbounded ρ-functions are highly sensitive to outlying residuals [14].

Setting the gradient in (4) equal to zero, and letting $w_{it} := \frac{1}{2z}\frac{d\rho_{c_1}(z)}{dz}\big|_{z=\frac{r_{it}(\boldsymbol{\theta}_i)}{\hat{\sigma}_{(i)}}}$ and $[\mathbf{W}_i]_{tt} = w_{it}$ for the corresponding diagonal matrix of weights, one obtains

$$(\boldsymbol{\Phi}_{\setminus ij}^\top \mathbf{W}_i \boldsymbol{\Phi}_{\setminus ij} + \lambda \mathbf{I})\boldsymbol{\theta}_i = \boldsymbol{\Phi}_{\setminus ij}^\top \mathbf{W}_i \mathbf{x}_i \tag{5}$$

where $\boldsymbol{\Phi}_{\setminus ij} := [\phi(\boldsymbol{\chi}_{\setminus ij}[1]), \ldots, \phi(\boldsymbol{\chi}_{\setminus ij}[T])]^\top$. Notice that observations with larger residuals receive lower weights as w_{it} is a decreasing function of $|r_{it}(\boldsymbol{\theta}_i)|$. As the weights are a function of $\boldsymbol{\theta}_i$ and vice versa, an alternating minimization scheme for solving (4) naturally arises. In particular, for a fixed $\boldsymbol{\theta}_i$, the resulting residuals and thus the corresponding matrix $\mathbf{W}_i$ are computed. With $\mathbf{W}_i$ now kept fixed, the estimate of $\boldsymbol{\theta}_i$ is updated as per (5). This alternating scheme can be shown to descend at each iteration [13].

Even though a bounded ρ-function enables a highly robust estimator, the price paid is nonconvexity in (4). Fortunately, it can be shown that with a high BDP and strongly consistent initialization, every local optimum will correspond to an estimator with high BDP and the same efficiency as the global optimum [25, 23]. To obtain such an initial estimate, the non-robust sum of squared residuals in (3) is replaced by a squared robust estimator of scale (cf. (2)) yielding

$$\hat{\boldsymbol{\theta}}_i^{(0)} \in \underset{\boldsymbol{\theta}_i \in \mathbb{R}^{2m}}{\arg\min} \quad T\hat{\sigma}^2_{\rho_{c_0}}(\mathbf{r}_i(\boldsymbol{\theta}_i)) + \lambda \|\boldsymbol{\theta}_i\|_2^2 \tag{6}$$

where $c_0 < c_1$ is selected so as to ensure consistency of $\hat{\sigma}_{\rho_{c_0}}(\cdot)$ at normal data; see also [13]. Equating the gradient to zero, it can be seen that the stationary points of (6) satisfy an equation similar to (5), albeit with $[\mathbf{W}_i]_{tt} = 2w_{it}$, and $\lambda \to (\lambda/T)\sum_{\tau=1}^T 2w_{i\tau}(r_{i\tau}(\boldsymbol{\theta})/\hat{\sigma}_{\rho_{c_0}}(\mathbf{r}_i(\boldsymbol{\theta}_i)))^2$. An alternating algorithm similar to that employed for solving (4) is thus leveraged here as well [13, 23]. It can be shown that for $\delta = 0.5[1 - \frac{2m}{T}]$ it holds that $\text{BDP}(\hat{\boldsymbol{\theta}}_i) \geq 0.5[1 - \frac{2m}{T}]$, even as $\lambda \to 0$; see e.g., [16, Thm. 5.8]. Moreover, strong consistency of $\hat{\boldsymbol{\theta}}_i$, under relatively mild assumptions, follows from [23, Thm. 3].

Remark 2. A question that naturally arises is why $\hat{\boldsymbol{\theta}}_i^{(0)}$ is not adopted as the final estimate, altogether bypassing (4). It turns out that the efficiency of $\hat{\boldsymbol{\theta}}_i^{(0)}$ in (6) at normal data as $\lambda \to 0$ and for a desired asymptotic BDP of 0.5, cannot exceed 0.33, which is extremely low [8]. Combining a high BDP estimator (cf. (6)) with a highly efficient one (cf. (4)) leads to a scheme that enjoys both properties [25].

As outlying observations may yield outlying residuals, and the correlation coefficient is known to be highly sensitive to outliers [6, 22], here we will explore robust alternatives. To begin, note that the population correlation coefficient for a pair of random variables (RVs) can be written solely as a function of the variances of properly chosen surrogate RVs. For the arbitrary pair X_i, X_j with corresponding means μ_i, μ_j and variances v_i^2, v_j^2, the population correlation coefficient can be expressed as

$$\varrho_{X_i,X_j} = \frac{\mathbb{V}(Z_{ij}) - \mathbb{V}(Z'_{ij})}{\mathbb{V}(Z_{ij}) + \mathbb{V}(Z'_{ij})} \tag{7}$$

where $Z_{ij} := \frac{X_i-\mu_i}{v_i} + \frac{X_j-\mu_j}{v_j}$, $Z'_{ij} := \frac{X_i-\mu_i}{v_i} - \frac{X_j-\mu_j}{v_j}$ and $\mathbb{V}(Z_{ij})$ denotes the variance of Z_{ij} [6]. A robust equivalent of the sample correlation coefficient can be obtained by replacing the sample variance with a robust counterpart thereof [22]. A natural choice is the squared Q_n estimator of dispersion, as it does not require centering and it is endowed both with a BDP of 0.5, which is the maximum attainable for scale equivariant estimators, and with very high efficiency [20]. For an arbitrary vector $\mathbf{r} \in \mathbb{R}^T$, the Q_n estimator is obtained as $\hat{v}_{Q_n}(\mathbf{r}) = 2.219\{|r_k - r_l|; k < l, k \in [T], l \in [T]\}_{(p)}$, where $p = \binom{\lfloor T/2 \rfloor + 1}{2}$ and $\mathcal{S}_{(p)}$ denotes the p-th smallest element of the set $\mathcal{S}$. The robust sample PC coefficient of $\mathbf{x}_i, \mathbf{x}_j$ is finally obtained by replacing $\mathbb{V}(Z_{ij})$ in (7) with $\hat{v}^2_{Q_n}(\hat{\mathbf{z}}_{ij})$, where $\hat{\mathbf{z}}_{ij} = (1/\hat{v}_{Q_n}(\mathbf{r}_i(\hat{\boldsymbol{\theta}}_i)))[\mathbf{r}_i(\hat{\boldsymbol{\theta}}_i) - \text{med}(\mathbf{r}_i(\hat{\boldsymbol{\theta}}_i))\mathbf{1}] + (1/\hat{v}_{Q_n}(\mathbf{r}_j(\hat{\boldsymbol{\theta}}_j)))[\mathbf{r}_j(\hat{\boldsymbol{\theta}}_j) - \text{med}(\mathbf{r}_j(\hat{\boldsymbol{\theta}}_j))\mathbf{1}]$, with $\text{med}(\cdot)$ denoting the median, and similarly for $\mathbb{V}(Z'_{ij})$.

To summarize, for each pair of nodes (i, j) the regression coefficient vectors $\{\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\theta}}_j\}$ are obtained as per (4), thereby yielding the corresponding residual vectors $\{\mathbf{r}_i(\hat{\boldsymbol{\theta}}_i), \mathbf{r}_j(\hat{\boldsymbol{\theta}}_j)\}$. Our robust PC coefficient for the pair (i, j) is then obtained as the robust correlation coefficient between $\{\mathbf{r}_i(\hat{\boldsymbol{\theta}}_i), \mathbf{r}_j(\hat{\boldsymbol{\theta}}_j)\}$ using the robust sample correlation coefficient of the ensemble in (7).

4. NUMERICAL TESTS

In order to assess performance of the proposed approach, numerical tests were performed on synthetic and real data. As is customary, Huber's contamination model was used, where with probability (w.p.) $(1 - \varepsilon)$ observations come from the nominal model (inliers), and w.p. $\varepsilon \in [0, 0.5]$ they follow an outlier distribution [9]. Performance of the proposed approach was compared to that of linear PC and to that of a robust variant of graphical lasso (RGL) [24], for various contamination rates ε. Regarding the proposed approach, a radial basis function kernel with a width parameter of 10 was used, and $2m = 50$ spectral features were utilized. Regularization

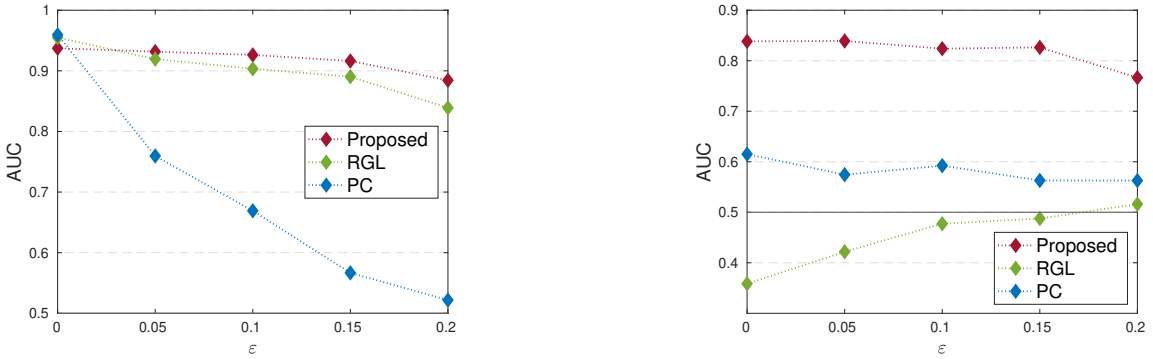

Fig. 1: AUCs for edge detection performance as a function of the contamination rate ε in the (a) linear and (b) nonlinear cases.

parameters for both RGL and the proposed approach were selected using five-fold cross-validation.

In the first two tests, synthetic data were used to allow for direct evaluation against the readily available ground truth graph structure. Random graphs comprising $|\mathcal{V}| = 20$ nodes were generated using the Barabási-Albert model [1]. The area under the receiver operating characteristics curve (AUC) was employed as figure of merit for detecting presence of edges. The presented results reflect the average performance across 20 realizations, each containing $T = 200$ observations.

In the first test, the nominal model was a multivariate Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{-1})$, where the precision matrix $\mathbf{\Omega}$ has the same sparsity pattern as the adjacency matrix of the corresponding graph. The outlier distribution was $\mathcal{N}(\mathbf{0}, 30\mathbf{I})$. This is a common choice, as outliers lie away from the central tendency, while the dependence structure that describes inliers is destroyed; see e.g., [7]. As depicted in Fig. 1a, all methods achieve excellent performance in the absence of outliers. This also demonstrates the normal efficiency of the proposed method. As ε increases though, the performance of PC deteriorates drastically, whereas performance of the proposed approach and RGL remains almost unaffected.

In the second test, *nonlinear* dependencies were present. Specifically, observations amounted to a nonlinear diffusion process over the graph. First, observations for the highest-degree node $\nu_1 \in \mathcal{V}$ were drawn from the uniform distribution $\mathcal{U}[0, 2]$. Observations at one-hop neighbors of ν_1, let $j \in \mathfrak{N}(\nu_1)$, were obtained as $x_j[t] = x_{\nu_1}^2[t] + u_j[t]$, where $u_j \sim \mathcal{U}[0, 0.1]$. With observations $\{x_j[t]; j \in \mathfrak{N}(\nu_1)\}$ kept fixed, the same noisy quadratic model was applied until the leaf nodes of the graph were reached. The outlier distribution was $x_i[t] \sim \mathcal{U}[0, 1] \quad \forall i \in \mathcal{V}$. As illustrated in Fig. 1b, even in the absence of outliers, PC and RGL perform poorly, with their AUCs being close to 0.5. The proposed approach on the other hand, achieves high estimation accuracy in the absence of outliers, and it is also robust to outliers, as evidenced by the small loss in performance as ε increases.

In the third test, real data comprising closing prices of $|\mathcal{V}| = 50$ stocks from the S&P index on all trading days in the years 2003-2007 were sourced from "Oxford-Man Institute's realized library" [21]. As no ground truth graph connectivity is available, the robustness of each approach to outliers was assessed using $\|\hat{\boldsymbol{\varrho}}_\varepsilon^{(M)} - \hat{\boldsymbol{\varrho}}_0^{(M)}\|_F / \|\hat{\boldsymbol{\varrho}}_0^{(M)}\|_F$, where $\hat{\boldsymbol{\varrho}}_\varepsilon^{(M)}$ is the matrix of PC coefficients estimated by method M at contamination rate ε, and $\hat{\boldsymbol{\varrho}}_0^{(M)}$ denotes the corresponding matrix estimated at uncontaminated data. Outliers were drawn from $\mathcal{N}(\mathbf{0}, 3\mathbf{I})$, and replaced actual observations. As evidenced in Table 1, the plain PC estimate is altered drastically relative to the uncontaminated case, even at the lowest contamination level ($\varepsilon = 0.05$), whereas the effect of outliers on the graph structure estimates obtained by RGL and the proposed approach remained limited, even at higher values of ε.

Table 1: Robustness in the S&P stocks dataset.

	ε			
	0.05	0.1	0.15	0.2
PC	0.729	0.793	0.820	0.844
RGL	0.059	0.097	0.126	0.156
Proposed	0.067	0.095	0.116	0.148

5. CONCLUSIONS

The present work introduced a novel approach to identifying graph connectivity using partial correlations. Highly robust and efficient bisquare M-estimators were utilized to learn nonlinear functions, scalability was effected through the use of spectral Fourier features, and a robust association measure was used to correctly assess correlations between fit residuals. The result is a robust approach to accurately estimating graph connectivity even when nonlinear mediation effects and outliers are present. Numerical tests demonstrated the benefits of the novel approach relative to existing alternatives.

6. REFERENCES

[1] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[2] X. Dong, D. Thanou, M. Rabbat, and P. Frossard, "Learning graphs from data: A signal representation perspective," *IEEE Signal Processing Magazine*, vol. 36, no. 3, pp. 44–63, 2019.

[3] M. Finegold and M. Drton, "Robust graphical modeling of gene networks using classical and alternative t-distributions," *The Annals of Applied Statistics*, vol. 5, no. 2A, pp. 1057 – 1080, 2011.

[4] J. Gao, B. Barzel, and A.-L. Barabási, "Universal resilience patterns in complex networks," *Nature*, vol. 530, no. 7590, pp. 307–312, 2016.

[5] G. B. Giannakis, Y. Shen, and G. V. Karanikolas, "Topology identification and learning over graphs: Accounting for nonlinearities and dynamics," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 787–807, 2018.

[6] R. Gnanadesikan and J. R. Kettenring, "Robust estimates, residuals, and outlier detection with multiresponse data," *Biometrics*, vol. 28, no. 1, pp. 81–124, 1972.

[7] K. Hirose, H. Fujisawa, and J. Sese, "Robust sparse Gaussian graphical modeling," *Journal of Multivariate Analysis*, vol. 161, pp. 172–190, 2017.

[8] O. Hössjer, "On the optimality of S-estimators," *Statistics & probability letters*, vol. 14, no. 5, pp. 413–419, 1992.

[9] P. J. Huber, "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73 – 101, 1964.

[10] G. Karanikolas, G. B. Giannakis, K. Slavakis, and R. M. Leahy, "Multi-kernel based nonlinear models for connectivity identification of brain networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6315–6319.

[11] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

[12] M. Lázaro-Gredilla, J. Quiñonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, "Sparse spectrum Gaussian process regression," *The Journal of Machine Learning Research*, vol. 11, pp. 1865–1881, 2010.

[13] R. A. Maronna, "Robust ridge regression for high-dimensional data," *Technometrics*, vol. 53, no. 1, pp. 44–53, 2011.

[14] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[15] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," *IEEE Signal Processing Magazine*, vol. 36, no. 3, pp. 16–43, 2019.

[16] L. Mili and C. W. Coakley, "Robust estimation in structured linear regression," *The Annals of Statistics*, vol. 24, no. 6, pp. 2593–2607, 1996.

[17] V. Öllerer, C. Croux, and A. Alfons, "The influence function of penalized regression estimators," *Statistics*, vol. 49, no. 4, pp. 741–765, 2015.

[18] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems*, 2008, pp. 1177–1184.

[19] S. H. Rao and G. L. Sievers, "A robust partial correlation measure," *Journal of Nonparametric Statistics*, vol. 5, no. 1, pp. 1–20, 1995.

[20] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273–1283, 1993.

[21] N. Shephard and K. Sheppard, "Realising the future: forecasting with high-frequency-based volatility (heavy) models," *Journal of Applied Econometrics*, vol. 25, no. 2, pp. 197–231, 2010.

[22] G. Shevlyakov and P. Smirnov, "Robust estimation of the correlation coefficient: An attempt of survey," *Austrian Journal of Statistics*, vol. 40, no. 1&2, pp. 147–156, 2011.

[23] E. Smucler and V. J. Yohai, "Robust and sparse estimators for linear regression models," *Computational Statistics & Data Analysis*, vol. 111, pp. 116–130, 2017.

[24] G. Tarr, S. Müller, and N. C. Weber, "Robust estimation of precision matrices under cellwise contamination," *Computational Statistics & Data Analysis*, vol. 93, pp. 404–420, 2016.

[25] V. J. Yohai, "High breakdown-point and high efficiency robust estimates for regression," *The Annals of Statistics*, pp. 642–656, 1987.

[26] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust Statistics for Signal Processing*. Cambridge University Press, 2018.