An Efficient Method for Identifying Statistical Interactors in Graphical Models

Alina Andrei, Ph.D.
Christina Kendziorski, Ph.D.
Title: An Efficient Method for Identifying Statistical Interactors in Graphical Models

Authors: Alina Andrei and C. Kendziorski*

Department of Biostatistics and Medical Informatics, 
University of Wisconsin, Madison, WI 53726

*Corresponding Author:

Christina Kendziorski

Department of Biostatistics and Medical Informatics

University of Wisconsin-Madison

6729 Medical Sciences Center

1300 University Avenue

Madison, WI 53703

Phone: (608) 262-3146

Fax: (608) 265-7916

Email: kendzior@biostat.wisc.edu
Summary:

Network reconstruction is a main goal of many biological endeavors. Graphical Gaussian models (GGMs) are often used since the underlying assumptions are well understood, the graph is readily estimated by calculating the partial correlation (paCor) matrix, and its interpretation is straightforward. In spite of these advantages, GGMs are limited in that interactions are not accommodated. As we show, when applied in the presence of interactions, the GGM framework can lead to incorrect inference regarding dependence. Identifying the exact dependence structure in this context is a difficult problem, largely because an analogue of the paCor matrix is not available and dependencies can involve many nodes. We here present a computationally efficient approach to identify bivariate interactions in networks. A key element is recognizing that interactions have a marginal linear effect, and as a result information about their presence can be obtained from the paCor matrix. Theoretical derivations for the exact effect are presented and used to motivate the approach; and simulations suggest that the method works well, even in fairly complicated scenarios. Practical advantages are demonstrated in analyses of data from a breast cancer study.

1 Introduction

The reconstruction of biological networks is a key component of efforts to better understand, diagnose, and treat disease. With advances in high throughput technologies, the data necessary for reconstruction are now available and so too are a variety of statistical reconstruction methods (Hartemink 2005; Werhli et al. 2006). Graphical Gaussian models (GGMs) are often used since the underlying assumptions are well understood, edges are easy to interpret, and estimation algorithms are readily available (for a review, see Li and Gui 2006). In short, nodes are assumed to follow a multivariate normal (MVN) distribution and an edge represents conditional dependence between
two nodes, given all other nodes (Lauritzen 1996; Whitaker 1990). The presence of edges corresponds exactly to non-zero partial correlations (paCors), and in this way estimation of the graph is equivalent to estimation of the paCor matrix. This simple, yet effective, framework has proven useful in a number of studies (Toh and Horimoto 2002; Wille et al. 2004; Matsuno et al. 2006; Ma et al. 2007; Keller et al. 2008).

In spite of the clear advantages, GGMs are limited in that they do not account for multiplicative dependence among nodes. Each variable is assumed normal, with conditional expectation a linear combination of some subset of the other variables. More precisely, a necessary and sufficient condition for a random vector \((X_1, X_2, \ldots, X_p)\) to follow a MVN distribution with mean \(\mu = (\mu_1, \ldots, \mu_p)\) and covariance matrix \(\Sigma\) is that for all \(k = 1, \ldots, p\), the conditional distribution of \(X_k\) given the remaining \(p - 1\) variables, \(X_{-k}\), depends on \(X_{-k}\) only through the conditional mean \(E[X_k|X_{-k}] = \alpha_k + B_kX_{-k}\), where \(B_k = \text{Cov}(X_k, X_{-k}) \cdot \text{Cov}(X_{-k})^{-1}\) and \(\alpha_k = \mu_k - B_k\mu_{-k}\) (Fisk, 1970). As a result, if \(X_k\) depends on \(X_i \cdot X_j\) (for \(i, j \neq k\)), for example, a GGM is not sufficient. This limitation is an important one, since interactions are common in biological studies of individual complex traits (Shiozawa et al., 2000; Wittenburg et al., 2001; Carlberg et al., 2003; Hanlon et al., 2006), studies of groups of metabolites (Wentzell et al., 2007), and high-throughput studies of gene expression (Kirst et al., 2005; Brem and Kruglyak 2005; West et al., 2007).

In practice, one often proceeds with GGM estimation and inference in the presence of interactions. As we demonstrate in Section 2, doing so can lead to incorrect inference regarding dependence. Inferring the correct dependence structure is challenging when interactions are present, since an analogue of the paCor matrix that completely determines the dependence structure is not available, and dependencies can involve many nodes. The difficulty is exacerbated for studies with a large number of nodes \((p)\) and relatively small sample size \((n)\).

Cox and Wermuth (1994) recognized this problem. In an investigation of the possible reasons
for departures from MVN, they focused on the presence of pairwise interactions, noting concern “that the reduction of the observations to covariance matrices overlooks important features of the dependences of intrinsic interest”. They proposed an approach to identify the interactions that evaluates all possible linear regression models among the variables considered allowing for single pairwise interaction terms. Interactions with large t-statistics are identified as important and considered for further empirical investigation. Although useful when few nodes are available, the approach cannot be easily scaled to cases with large $p$, and is not possible when $n < p$.

As in Cox and Wermuth (1994), we too consider the special case of non-MVN where the full conditional distributions are Gaussian, but the joint is not, due to pairwise interactions. Within this context, we propose a computationally efficient approach that uses GGM estimation techniques to identify variables affected by interactions, and thereby identify regions in a resulting graph where edges may not imply direct or linear dependencies. Our approach is presented in Section 3, with much of the theoretical support relegated to the Appendix. The method relies on estimation of multiple paCor matrices, and as a result is only slightly more computationally demanding than traditional GGM reconstruction. Simulation studies in Section 4 show that the approach identifies interactors with high sensitivity for practical levels of heritability and moderate sample sizes, including cases where $n < p$. Further evaluation is given in Section 5, where we identify novel interactors using data from a study of breast cancer.


2 A Simple Motivating Example

Let $X = (X_1, \ldots, X_p)$ denote $p$ random variables. The paCor of $X_i$ and $X_j$ is defined as the correlation of the residuals obtained after linearly regressing each of the two on the remaining $p - 2$ variables. In the case of GGMs, where $X$ is MVN, the paCor matrix completely specifies the dependence structure among the variables: the paCor between two nodes is zero if and only if the two nodes are conditionally independent given all other variables. If $X$ is not MVN, this interpretation of the paCor matrix does not hold.

An interesting, well studied, and practically relevant case of non-MVN concerns the situation in which all full conditional distributions are normal, but the joint is not MVN, due to pairwise interactions. Consider two simple examples where $X_1$, $X_2$, and $\epsilon$ are independent normals (more general examples are provided in the Appendix - see Remark 1). For the first case, let $p = 3$ and $X_3 = a_1 X_1 + a_2 X_2 + b X_1 \cdot X_2 + \epsilon$. The $\text{paCor}(X_3, X_1|X_2) = \frac{\sqrt{\text{var}(X_1)(a_1+bE(X_2))}}{\sqrt{b^2 \text{var}(X_1)\text{var}(X_2) + \text{var}(X_2)(a_1+bE(X_2))^2+\text{var}(\epsilon)}}$, which is zero if $a_1+bE(X_2) = 0$. In this case, the edge between $X_1$ and $X_3$ is not detected in a GGM, even though the two are conditionally dependent.

The precise constraint on the parameters may ensure that this case rarely happens in practice. For the second example, let $p = 4$ and $X_k = a_{1k} X_1 + a_{2k} X_2 + b_k X_1 \cdot X_2 + \epsilon$ for $k = 3, 4$. The $\text{paCor}(X_3, X_4|X_1, X_2) = \frac{b_3 b_4 \text{var}(X_1)\text{var}(X_3)}{\sqrt{b_3^2 \text{var}(X_1)\text{var}(X_2)+\sigma_3^2} \sqrt{b_4^2 \text{var}(X_1)\text{var}(X_2)+\sigma_4^2}}$, which is non-zero for any non-zero $b_3$ and $b_4$. In other words, an edge would be detected between $X_3$ and $X_4$ even though they are conditionally independent given $X_1$ and $X_2$. In short, the one-to-one relationship between non-zero paCors and conditional dependencies that exists in a GGM breaks down when multiplicative effects are present. Frequently, incorrect inference will involve declaring a pair of variables conditionally dependent when in fact they are not.
3 Methods

As in the case of GGM reconstruction, our method is also based on paCor estimates. A key element is recognizing that interactions have a marginal linear effect, and as a result, information about their presence can be obtained from the paCor matrix. In short, the first step consists of identifying nodes that are potentially affected by interactors; these are the nodes that have at least two non-zero paCor coefficients (one corresponding to each interactor). An appended matrix is then constructed for each such node, consisting of its significant associations (detected in the first step) as well as all possible pairwise interactions. Partial correlations are then re-estimated and used to identify true interactors. The construction of the appended matrix is supported by noting that nodes affected by an interaction have a paCor with the variables in the interaction that directly depends on the size of the interaction coefficient. Consider the first example in the previous section where \( X_3 = a_1 X_1 + a_2 X_2 + bX_1 \cdot X_2 + \epsilon \). Then \( \text{paCor}(X_3, X_1|X_2) \) and \( \text{paCor}(X_3, X_2|X_1) \) are non-zero (except in very particular cases) and so \( X_3 \) would be a node identified in step 1 as potentially affected by an interaction and an appended matrix for \( X_3 \) would be constructed with columns \( \{X_1, X_2, X_1 \cdot X_2\} \). The construction of this matrix is motivated by noting that \( \text{paCor}(X_3, X_1 \cdot X_2|X_1, X_2) \) depends only on the coefficient of the interactor (a more general claim is below). The detailed steps of the proposed approach follow for a more general setting.

Description of the proposed method

Step 1. For variables \( X_1, \ldots, X_p \), identify the non-zero paCor coefficients. A number of methods are now available for estimating the paCor matrix for sparse graphs when \( n << p \) (Schäfer and Strimmer 2005a, 2005b; Li and Gui 2006; Meinshausen and Bühlman 2006; Yuan and Lin 2007; Friedman et al. 2007). We rely on the method of Schäfer and Strimmer (2005a, 2005b), as it provides an estimated paCor matrix (using the optimal shrinkage estimator of Ledoit and Wolf.
as well as an empirical Bayes method to identify $\rho$Cors at a desired false discovery rate (FDR). The Schäfer and Strimmer approach is implemented in the R package GeneNet version 1.2.1, available from http://cran.r-project.org/doc/packages/.

**Step 2.** Consider the variables that have at least two non-zero entries in the $\rho$Cor matrix. Let $X_i$ denote such a variable and $X_{i_1}, \ldots, X_{i_k}$ ($k \geq 2$) the corresponding variables having non-zero $\rho$Cors with $X_i$. Form a new data matrix $M_i$ that includes these variables and all two-way interactors, namely:

$$X_i, X_{i_1}, \ldots, X_{i_k}, X_{i_1} \cdot X_{i_2}, \ldots, X_{i_{k-1}} \cdot X_{i_k}.$$

**Step 3.** Estimate the $\rho$Cor matrix of data $M_i$, for each variable $X_i$ associated with at least two other variables. If

$$|\rhoCor(X_i, X_{i_s} \cdot X_{i_{s+t}})| > \max(|\rhoCor(X_i, X_{i_s})|, |\rhoCor(X_i, X_{i_{s+t}})|)$$

for $s \leq k - 1, t \leq k - s$, then $X_{i_s} \cdot X_{i_{s+t}}$ is an interactor for $X_i$. This step is justified by the following claim:

**Claim:** Let $T = \{T_1, \ldots, T_n\}$ be a set of $n$ independent variables, $n \geq 2$, with $E(T_i) = \mu_i$ and $\sigma^2(T_i) = \sigma^2_i, i \in \{1, \ldots, n\}$ and assume that

$$Z = \sum_{i=1}^n a_i T_i + \sum_{1 \leq i < j \leq n} b_{ij} T_i \cdot T_j + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ and $\epsilon$ and $T_i$ are marginally independent. Let $\mathcal{Y} = \{T_1 \cdot T_2, T_1 \cdot T_3, \ldots, T_{n-1} \cdot T_n\}$ denote the set of all possible pairwise interactors. Then

$$|\rhoCor(Z, T_i \cdot T_j | T, \mathcal{Y}\setminus T_i \cdot T_j)| > \max(|\rhoCor(Z, T_i | T\setminus T_i, \mathcal{Y})|, |\rhoCor(Z, T_j | T\setminus T_j, \mathcal{Y})|)$$

if and only if $|b_{ij}| > \max \left( \frac{|a_i|}{\sqrt{1 + \sum_{k=1, k \neq i}^n \frac{\mu_k^2}{\sigma_k^2}}} \cdot \frac{|a_j|}{\sqrt{1 + \sum_{k=1, k \neq j}^n \frac{\mu_k^2}{\sigma_k^2}}} \right), (*)$

A detailed proof of this result is provided in the Appendix. The claim indicates that our approach can be used to identify interactors when the effect $b_{ij}$ is large enough, as indicated by the above inequality. We shown in Section 5 that this restriction is not prohibitive, as a number of interactors are identified in the case study considered.
**Step 4.** Draw the graph. The resulting graph consists of the nodes (in yellow; isolated nodes are not shown) and edges identified in step 1. Additional nodes (in purple) and edges are added to highlight the interactors identified in Step 3.

### 4 Simulation Studies

In order to investigate the performance of the proposed approach, we have conducted computer simulations for networks of varying degrees of sparseness, heritability, and sample sizes. Each simulation consists of 100 nodes of which $100 - N$ are independent Gaussians $X_1 \sim N(\mu_1, \sigma_1^2), \ldots, X_{100-N} \sim N(\mu_{100-N}, \sigma_{100-N}^2)$, while the remaining $N$ nodes are generated according to the models:

\[
Y_i \sim N(\delta X_i + \xi X_{i+1}, \sigma^2), i \in \{1, \ldots, m\},
\]

\[
Y_i \sim N(\alpha X_i + \beta X_{i+1} + \gamma X_i \cdot X_{i+1}, \sigma^2), i \in \{m + 1, \ldots, 100 - N\}.
\]

Simulation sets I, II, and III correspond to $N = 15$, $N = 20$, and $N = 30$, respectively, with $m = 5$, $m = 10$, and $m = 15$. In genetics studies, heritability of a trait is an important measure representing the proportion of phenotypic variation in a population that is attributable to genetic variation among individuals. To specify a heritability for each simulated data set, we set $\sigma_\epsilon = 1$, $\sigma'_s = 1$, and choose appropriate values for the remaining parameters, according to the following:

\[
\text{Heritability}(Y_i) = \frac{\alpha^2 \sigma_i^2 + (\beta + \alpha \mu_{i+1})^2 \sigma_{i+1}^2 + (\gamma + \alpha \mu_i)^2 \sigma_i^2}{\alpha^2 \sigma_i^2 + (\beta + \alpha \mu_{i+1})^2 \sigma_{i+1}^2 + (\gamma + \alpha \mu_i)^2 \sigma_i^2 + \sigma^2}, i \in \{1, \ldots, m\},
\]

\[
\text{Heritability}(Y_i) = \frac{\delta^2 \sigma_{i+6}^2 + \xi^2 \sigma_{i+50}^2}{\delta^2 \sigma_{i+6}^2 + \xi^2 \sigma_{i+50}^2 + \sigma^2}, i \in \{m + 1, \ldots, 100 - N\}.
\]

Sample sizes from 50 to 250 were considered for heritabilities ranging from 40% to 80%. These values are in the range considered in many genomic studies (Zeng 1994; Brem and Kruglyak 2005).
2005). We generated 1000 data sets under each simulation scenario, heritability level, and sample size considered.

The proposed approach was applied to each simulated data set using an FDR threshold in step 1 of 10%. Figure 1 shows the average power, which exceeds 60% for heritability levels and sample sizes exceeding 0.6 and 100, respectively. The power for detecting main effects and interactors is reasonable for lower heritability levels, particularly as the sample size increases. Average FDR was < 0.1, and is therefore not shown.
Figure 1: Three simulation scenarios are considered (upper, middle, and bottom panels). In each scenario, 1000 data sets were generated for each sample size (x-axis) and each level of heritability (varying color). Average power to detect main effects (left) and interactions (right) is shown.
5 Case Study: Breast Cancer Data

We applied our approach to microarray data from the Duke Breast Cancer SPORE frozen tissue bank (West et al., 2001), available at http://data.cgt.duke.edu/WEST/PNASCel.zip. The data consist of 49 tissue samples and 7129 expression measurements, recorded using Affymetrix hu6800 chips. Pre-processing was done using Robust Multi-array Average (RMA) (Irizarry et al. 2003) to obtain a single, normalized score of expression for each probe set on each array. When multiple probe sets were used for the same gene, we selected the probe set measurement with the highest RMA average across the 49 samples.

Because it is difficult to visualize, interpret, and understand a network constructed from over 7000 nodes, we chose to focus on networks constructed around genes that are likely involved in cancer. One of these genes is PTPRE, a member of the protein-tyrosine phosphatases gene family (PTPs). PTPs are known to control tyrosine phosphorylation, an important signaling mechanism in eukaryotic cells. Moreover, recent evidence has shown that PTPs can function as tumor suppressors in different types of human cancers (Östman et al. 2006). In an attempt to detect direct interactions among genes that might be involved in similar biological pathways, we considered a set of 173 genes that are correlated with PTPRE at level 0.7 or higher.

The resulting network is shown in Figure 2 and consists of 128 edges, of which 16 represent interacting genes. In order to control the proportion of possibly spurious findings, the FDR has been controlled at 0.1, as in the simulation studies. By accounting for interactors, the reconstructed network provides additional insight into the nature of the direct associations (additive or multiplicative effects) that are detected.
Figure 2: Network reconstructed from 173 genes correlated with gene PTPRE. A total of 128 edges were obtained. Unconnected nodes are not shown. Linear effects are shown in yellow; interactors are shown in purple.
Figure 3: Panel (a): Residuals of the additive model for ZNF76 versus PLCD1*ELL2. Panel (b): Residuals of the additive model for RPE65 versus PIGR*ALKBH1.

Consider, for example, the subgraph around ZNF76 (zinc finger protein 76), an important gene that functions as a transcriptional repressor through its interaction with the known TATA-binding protein (TBP) (Zheng and Yang 2004). Our results suggest that ZNF76 depends directly on PDX1, MYBPC2, PLCD1, ELL2, and the interactor PLCD1-ELL2. Recent studies show that PDX1, the pancreatic duodenal homeobox-1 gene, is an early marker for several types of human cancers, including breast cancer, and could potentially be used as a diagnostic parameter (Wang et al. 2005). In addition, PLCD1, phospholipase C delta 1, is a protein coding human gene, found to play an important suppressive role in the development and progression of esophageal cell carcinoma (Fu et al. 2007).

As a way to further investigate this dependence, we fitted linear regression models with and without the interaction term. The interactor was highly significant, with an absolute effect size larger than 1 (SE = 0.15; p-value < 0.001). Moreover, the interaction explains a substantial
proportion of the heritability for ZNF76: the additive effects alone account for 34\% of the genetic variance, which increases to 73\% when the interaction term is included. The presence of the interactor in the data is further reinforced in panel (a) of Figure 3, which shows a non-random pattern when plotting the residuals of the additive model against the interaction of PLCD1 and ELL2. Finally, the interaction model is favored by the Bayesian Information Criterion (BIC). These types of results also hold for most of the other interactors identified. The non-random scatter shown in Figure 3, panel (b), emphasizes the dependence of the gene RPE65 (center of the network) on the interactor PIGR*ALKBH1. Overall, 11 of the 16 interactors identified in the full network were found significant based on local linear regression models (p-value < 0.05), the interaction models resulted in comparably lower BICs, andheritabilities were significantly increased when accounting for the interactors.

6 Discussion

Properly reconstructed biological networks have the potential to elucidate many important questions in biology and, with recent technological advances, the measurements required for reconstruction are now largely available. As a result, network reconstruction efforts abound. The GGM framework is often used since computation and interpretation are straightforward. However, GGMs do not allow for interacting nodes due to the underlying MVN assumption. As we show here for two simple cases, proceeding with GGM estimation and interpretation in the presence of interactions can result in both false positive and false negative identifications of conditional dependence.

Identifying the underlying graph structure in the presence of interactors is a challenging problem. We have considered the specific case where nodes are conditionally normal, but the
joint is not MVN due to the presence of interactions. The form of the joint distribution for this case has been characterized in the bivariate (Bhattacharya 1943) and multivariate settings (Arnold 1999). In spite of this, an analogue of the paCor matrix that uniquely determines linear and higher order dependencies does not exist. Deriving such a measure for this family of densities is a challenging open question. In the meantime, the proposed approach should prove useful in identifying interactions and constructing informative graphs.

The graph derived from the Duke Breast Cancer study provides an example, with 16 pairwise interactions identified from the 173 nodes considered. The graph reveals information regarding direct dependencies, and perhaps more importantly highlights variables that warrant further investigation, namely those involving interactions. We fit local linear regression models to more carefully investigate the dependence among such variables. The regression checks are clearly limited since they are local, only considering relatively few nodes at a time; and they cannot identify interactors with small effects (where small is defined as in Section 3). In spite of the limitations, local regression analyses can be informative, as they provide a straightforward test for the presence of interactions, can be used to assess the impact of the interactions by calculating the percent variance explained, can be used to directly compare models with and without an interaction term (e.g. via BICs), and can provide insight into possibly spurious edges. In the Duke study, 69% of the interactors identified were supported by a local regression analysis and in many cases the percent variance explained was more than doubled when interaction terms were included. BICs were also improved. Although we did not find evidence of multiple nodes affected by a common interactor (such as in the second example provided in Section 2), local regressions could be used in such cases in an effort to identify false edges in the graph.

In practice, the graph identified depends on a number of factors, and it is worthwhile to investigate how moderate variations impact the results. An obvious value to consider is the FDR
threshold, which in part determines the number of candidate interactors detected in the first step of the approach. We considered an FDR threshold of 0.1 in the simulations and case study. However, it could certainly be the case that for some datasets a threshold of 0.1 will yield an excess of edges and interactors that are difficult to interpret. Alternatively, there could be datasets for which a slightly higher FDR cutoff is required. We recommend varying the FDR threshold and checking that edges calculated at the more conservative threshold are largely preserved when the threshold is increased. That was the case for the Duke study presented here (results not shown).

Another factor to consider is whether or not variables are scaled. It is well known that predictors that are additive on one scale may exhibit an interaction when a different scale is used, and conversely (Schork and Frankel 1996, Greenland and Rothman 1998, Cordell 2002). In the context of our method, some of the coefficients important for identification of interactions (step 3 in Methods) are not invariant to scaling, and as a result, the interactors identified on one scale might differ from those obtained on another. We here did not scale the variables, noting as in Schork and Frankel (1996) that doing so may facilitate the identification of interactors.

The number of edges and interactions detected will also depend on the number of nodes initially considered, and on the method used for node selection. Our approach relies on estimation of a paCor matrix, and since good approaches now exist for estimation in the case of \( n < p \), the number of nodes considered is no longer restricted by the sample size. Nevertheless, a choice needs to be made, and the method used for feature selection will impact the number of interactors identified. In our example, we chose to focus on sets of transcripts that show a high correlation with a gene known to be associated with cancer. Alternative approaches could be to consider all nodes within a particular known pathway, or to consider genes correlated with some clinical outcome of interest.

In summary, we have presented an approach to identify statistical interactions in networks. The main advantages compared with other network reconstruction approaches that allow for non-linear
relationships among nodes (e.g. Friedman et al. 2000) are that the approach makes relatively few, but testable, assumptions, does not rely on prior specification, and is computationally efficient, requiring little increase in computation over traditional GGM fitting. Extensions to more general networks are underway and should prove useful particularly for network reconstructions involving continuous phenotype and discrete genotype data.

7 Appendix

We here provide a detailed proof of the claim mentioned in Section 3. The proof assumes that the sample size is larger than the number of variables.

**Main Claim:** Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be a set of $n$ independent variables, $n \geq 2$, with $E(X_i) = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2$, $i \in \{1, \ldots, n\}$ and assume that $Z = \sum_{i=1}^n a_i X_i + \sum_{1 \leq i < j \leq n} b_{ij} X_i \cdot X_j + \epsilon$, with $\epsilon \sim N(0, \sigma^2)$, independent from all $X_i$’s. Let $\mathcal{Y} = \{Y_{ij} = X_i \cdot X_j | 1 \leq i < j \leq n\}$ denote the set of all possible interactors. Then the following holds:

$$|\text{paCor}(Z, Y_{ij} | \mathcal{X}, \mathcal{Y} \setminus Y_{ij})| \geq \max(|\text{paCor}(Z, X_i | \mathcal{X} \setminus X_i, \mathcal{Y})|; |\text{paCor}(Z, X_j | \mathcal{X} \setminus X_j, \mathcal{Y})|) \quad (*)$$

if and only if $|b_{ij}| \geq \max(\frac{|a_i|}{\sqrt{1 + \sum_{k=1, k \neq i}^n \frac{\mu_k^2}{\sigma_k^2}}}; \frac{|a_j|}{\sqrt{1 + \sum_{k=1, k \neq j}^n \frac{\mu_k^2}{\sigma_k^2}}})$.

In particular, if $|b_{ij}| \geq \max(|a_i|; |a_j|)$, then inequality $(*)$ is satisfied.

**Proof:** We compute explicit forms for $|\text{paCor}(Z, Y_{ij} | \mathcal{X}, \mathcal{Y} \setminus Y_{ij})|$, $|\text{paCor}(Z, X_i | \mathcal{X} \setminus X_i, \mathcal{Y})|$, $|\text{paCor}(Z, X_j | \mathcal{X} \setminus X_j, \mathcal{Y})|$. To simplify computations, assume first that $\sigma_i = 1$, for all $i$. The general claim follows easily by rescaling the variables. The proof uses the following recursive formula (see Kendall’s, Advanced Theory of Statistics, and for a detailed proof, Pearson (1916)): if $T_1, \ldots, T_n$ are random variables, then the paCor of $T_1$ and $T_2$ with respect to the remaining $n - 2$
variables can be computed recurrently by:
\[
\text{paCor}(T_1, T_2 | T_3, ..., T_n) = \frac{\text{paCor}(T_1, T_2 | T_3, ..., T_{n-1}) - \text{paCor}(T_1, T_n | T_3, ..., T_{n-1}) \text{paCor}(T_2, T_n | T_3, ..., T_{n-1})}{\sqrt{1 - \text{paCor}^2(T_1, T_n | T_3, ..., T_{n-1})} \sqrt{1 - \text{paCor}^2(T_2, T_n | T_3, ..., T_{n-1})}}.
\]

This formula does not depend on the underlying distribution. Note also that for \( n = 3 \), the paCor can be expressed in terms of the correlation as follows:
\[
\text{paCor}(T_1, T_2 | T_3) = \frac{c(T_1, T_2) - c(T_1, T_3) c(T_2, T_3)}{\sqrt{1 - c^2(T_1, T_3)} \sqrt{1 - c^2(T_2, T_3)}}.
\]

We show first the following intermediary claims:

\textbf{Claim 1:} \( \text{paCor}(Z, X_i | \mathcal{X} \setminus X_i) = \frac{c(Z, X_i)}{\sqrt{1 - \sum_{j=1, j \neq i}^n c^2(X_j, Z)}.} \)

\textbf{Proof of Claim 1:} Without loss of generality, one may assume that \( i = 1 \). Note that \( \text{paCor}(Z, X_1 | X_2) = \frac{c(Z, X_1)}{\sqrt{1 - c^2(Z, X_2)}} \) and that \( \text{paCor}(X_1, X_j | \mathcal{X} \setminus \{X_1, X_j\}) = 0, (\forall) j \), by the independence assumption. The claim follows easily by using repeatedly the recursive formula.

\textbf{Remark 1.} Due to the presence of interactors, the partial correlation of two variables can be zero even if they are directly dependent. In our setting, \( \text{paCor}(Z, X_i | \mathcal{X} \setminus X_i) = 0 \) if \( c(Z, X_1) = 0 \), i.e., \( a_i + \sum_{j \neq i}^n b_{ij} \mu_j = 0 \). This can affect step 1, where non-zero paCors are identified. However, the very specific coefficient configuration likely ensures that this does not happen very often.

\textbf{Claim 2:} \( \text{paCor}(Z, Y_{ij} | X_1, ..., X_m) = \frac{b_{ij}}{\sigma_Z \sqrt{1 - \sum_{i=1}^m c^2(X_i, Z)}}, (\forall) 2 \leq m \leq n. \)

\textbf{Proof of Claim 2:} One may assume that \( i = 1 \) and \( j = 2 \). We proceed by induction on \( m \).

Consider first \( \text{paCor}(Z, Y_{12} | X_1, X_2) \). Let \( \sigma^2_{Y_{12}} \) denote the variance of \( Y_{12} \). The following formulas hold:
\[
c(X_1, Z) = \frac{(a_1 + \sum_{j=2}^n b_{ij} \mu_j)}{\sigma_Z},
\]
\[
c(X_1, Y_{12}) = \frac{\mu_2}{\sqrt{1 + \mu_2^2 + \mu_1^2}}.
\]
\[ c(X_2, Z) = \frac{(a_2 + \sum_{j=3}^n b_{2j} \mu_j + b_{12} \mu_1)}{\sigma_Z}. \]
\[ c(X_2, Y_{12}) = \frac{\mu_1}{\sqrt{1 + \mu_2^2 + \mu_1^2}}. \]
\[ c(Z, Y_{12}) = \frac{a_1 \mu_2 + a_2 \mu_1 + b_{12}(1 + \mu_1^2 + \mu_2^2) + \sum_{j=3}^n b_{1j} \mu_j \mu_2 + \sum_{j=3}^n b_{2j} \mu_1 \mu_j}{\sigma_Z \sigma_{Y_{12}}}. \]

\[ \text{paCor}(Z, Y_{12}|X_1) = \frac{(a_2 b_{12} + b_{12}(1 + \mu_1^2) + \sum_{j=3}^n b_{2j} \mu_1 \mu_j)}{\sqrt{\sigma_Z^2 - (a_1 + \sum_{j=2}^n b_{1j} \mu_j)^2 \sqrt{1 + \mu_1^2}}}. \]
\[ \text{paCor}(X_2, Z|X_1) = \frac{(a_2 + b_{12} \mu_1 + \sum_{j=3}^n b_{2j} \mu_j)}{\sqrt{\sigma_Z^2 - (a_1 + \sum_{j=2}^n b_{1j} \mu_j)^2}}. \]
\[ \text{paCor}(X_2, Y_{12}|X_1) = \frac{\mu_1}{\sqrt{1 + \mu_1^2}}. \]

The above expressions and the recursive formula give
\[ \text{paCor}(Z, Y_{12}|X_1, X_2) = \frac{\text{paCor}(Z, Y_{12}|X_1) - \text{paCor}(Z, X_2|X_1) \text{paCor}(Y_{12}, X_2|X_1)}{\sqrt{1 - \text{paCor}(Z, X_2|X_1) \text{paCor}(Y_{12}, X_2|X_1)}} = \frac{b_{12}}{\sigma_Z \sqrt{1 - \sum_{i=1}^m c_i^2(X_i, Z)}}. \]

Assume that \( \text{paCor}(Z, Y_{12}|X_1, ..., X_{m-1}) = 0, \) (\( \forall \) \( m \geq 3 \)), and the recursive formula gives that \( \text{paCor}(Z, Y_{12}|X_1, ..., X_m) = \frac{\text{paCor}(Z, Y_{12}|X_1, ..., X_{m-1})}{\sqrt{1 - \text{paCor}(Z, X_m|X_1, ..., X_{m-1})}}. \) The conclusion follows by using claim 1.

**Claim 3:** \( \text{paCor}(Y_{ij}, Y_{mn}|X, Y \setminus \{Y_{ij}, Y_{mn}\}) = 0, \) (\( \forall \) \( (i, j) \neq (m, n) \)).

**Proof of Claim 3:** We show first that \( \text{paCor}(Y_{ij}, Y_{mn}|X) = 0. \) This is immediate if the interactors \( Y_{ij}, Y_{mn} \) do not share a variable. Next, assume that \( Y_{ij} = Y_{12} \) and \( Y_{mn} = Y_{13}. \) Then:
Claim 5: \( \text{paCor}(Y_{12}, Y_{13}|X_1) = \frac{c(Y_{12}, Y_{13}) - c(Y_{12}, X_1)c(Y_{13}, X_1)}{\sqrt{1 - c^2(Y_{12}, X_1)} \sqrt{1 - c^2(Y_{13}, X_1)}} = \frac{\mu_2 \mu_3}{\sigma_{Y_{12}} \sigma_{Y_{13}}} - \frac{\mu_2}{\sigma_{Y_{12}}} \cdot \frac{\mu_3}{\sigma_{Y_{13}}} = 0 \)

By applying the recursive formula repeatedly, we obtain that \( \text{paCor}(Y_{12}, Y_{13}|X, \mathcal{Y}\{Y_{12}, Y_{13}\}) = 0 \). Next, we show that \( \text{paCor}(Y_{12}, Y_{13}|X, \mathcal{Y}\{Y_{12}, Y_{13}\}) = 0 \). We have that:

\[
\text{paCor}(Y_{12}, Y_{13}|X, \mathcal{Y}|i,j) = \frac{\text{paCor}(Y_{12}, Y_{13}|X) - \text{paCor}(Y_{12}, Y_{14}|X) \text{paCor}(Y_{13}, Y_{14}|X)}{\sqrt{1 - \text{paCor}^2(Y_{12}, Y_{14}|X) \sqrt{1 - \text{paCor}^2(Y_{13}, Y_{14}|X)}}} = 0,
\]

and the claim follows easily by induction.

Claim 4: \( \text{paCor}(Z, Y_{ij}|\mathcal{X}, \mathcal{Y}|i,j) = \frac{b_{ij}}{\sqrt{b_{ij}^2 + \sigma_i^2}} \).

Proof of Claim 4: We show that \( \text{paCor}(Z, Y_{ij}|\mathcal{X}, \mathcal{Y}|i,j) = \frac{b_{ij}}{\sqrt{\sigma_Z^2 - \sum_{t=1}^n b_{ij} b_{st}^2}} \) (*), where \( \Delta = 1 - \sum_{i=1}^n c^2(X_i, Z) \). Again, assume that \( i = 1 \) and \( j = 2 \). From claim 3, it follows that \( \text{paCor}(Z, Y_{12}|\mathcal{X}, Y_{13}) = \frac{\text{paCor}(Z, Y_{12}|\mathcal{X})}{\sqrt{1 - \text{paCor}^2(Z, Y_{12}|\mathcal{X})}} \). From claim 2, we have that \( \text{paCor}(Z, Y_{13}|\mathcal{X}) = \frac{b_{13}}{\sigma_Z \sqrt{\Delta}} \). Then \( \text{paCor}(Z, Y_{12}|\mathcal{X}, Y_{13}) = \frac{b_{12}}{\sqrt{\sigma_Z^2 - b_{12}^2}} \) and identity (*) follows by using repeatedly the recursive formula. Note that \( \sigma_Z^2 = \sum_{i=1}^n (a_i + \sum_{j \neq i} b_{ij} \mu_j)^2 + \sum_{1 \leq i < j \leq n} b_{ij}^2 + \sigma_i^2 \), hence \( \sigma_Z^2 \Delta = \sum_{1 \leq i < j \leq n} b_{ij}^2 + \sigma_i^2 \) and the claim follows.

Next, we compute \( \text{paCor}(Z, X_i|\mathcal{X}\{X_i, \mathcal{Y}\}) \) and \( \text{paCor}(Z, X_j|\mathcal{X}\{X_j, \mathcal{Y}\}) \). Assume that \( i = 1 \) and \( j = 2 \).

Claim 5: \( \text{paCor}(Z, X_1|\mathcal{X}\{X_1, Y_{12}\}) = \frac{(a_1 + \sum_{j=3}^n b_{ij} \mu_j)}{\sqrt{1 + \mu_2^2 (\sum_{1 \leq i < j \leq n, \ (i,j) \neq (1,2)} b_{ij}^2 + \sigma_i^2) + (a_1 + \sum_{j=3}^n b_{ij} \mu_j)^2}} \).

Proof of Claim 5: From the recursive formula, \( \text{paCor}(Z, X_1|\mathcal{X}\{X_1, Y_{12}\}) = \frac{\text{paCor}(Z, X_1|\mathcal{X}\{X_1, Y_{12}\}) - \text{paCor}(Z, Y_{12}|\mathcal{X}\{X_1\}) \text{paCor}(X_1, Y_{12}|\mathcal{X}\{X_1\})}{\sqrt{1 - \text{paCor}^2(Z, X_1|\mathcal{X}\{X_1, Y_{12}\}) \sqrt{1 - \text{paCor}^2(X_1, Y_{12}|\mathcal{X}\{X_1\})}}} \) and so \( \text{paCor}(X_1, Y_{12}|X_2, X_3) = \frac{\text{paCor}(X_1, Y_{12}|X_2, X_3)}{\sqrt{1 - \text{paCor}^2(Y_{12}|X_2, X_3)}} \) since \( X_1, X_2, X_3 \) are independent. Similarly, we obtain that \( \text{paCor}(X_1, Y_{12}|\mathcal{X}\{X_1\}) = \text{paCor}(X_1, Y_{12}|X_2) = \frac{\mu_2}{\sqrt{1 + \mu_2^2}} \).
By using the recursive formula we obtain:

\[
paCor(Z, Y_{12}|X_2, \ldots, X_n) = \frac{\paCor(Z, Y_{12}|X_2)}{\sqrt{1 - \sum_{j=3}^{n} \paCor^2(Z, X_j|X_2)}} = \frac{c(Z, Y_{12}) - c(Y_{12}, X_2)c(Z, X_2)}{\sqrt{1 - c^2(Y_{12}, X_2)}} \cdot \frac{1}{\sqrt{1 - \sum_{j=3}^{n} c^2(Z, X_j)}}
\]

\[
= \frac{(a_1\mu_2 + b_{12} + \mu_2 \sum_{j=2}^{n} b_{1j}\mu_j)}{\sigma_Z \sqrt{1 + \mu_2^2}} \sqrt{1 - \sum_{j=2}^{n} c^2(Z, X_j)}.
\]

Hence

\[
paCor(Z, X_1|Y_{12}, \mathcal{X}\setminus X_1) = \frac{c(Z, X_1)}{\sqrt{1 - \sum_{j=2}^{n} c^2(Z, X_j)}} - \frac{(a_1\mu_2 + b_{12} + \mu_2 \sum_{j=2}^{n} b_{1j}\mu_j)}{\sigma_Z \sqrt{1 + \mu_2^2}} \sqrt{1 - \sum_{j=2}^{n} c^2(Z, X_j)} \cdot \frac{\mu_2}{\sqrt{1 + \mu_2^2}}
\]

\[
= \frac{a_1 + \sum_{j=3}^{n} b_{1j}\mu_j}{\sqrt{(1 + \mu_2^2)(\sum_{1\leq i<j\leq n, (i,j)\neq (1,2)} b_{ij}^2 + \sigma_i^2) + (a_1 + \sum_{j=3}^{n} b_{1j}\mu_j)^2}}.
\]

Next we account for all interactors containing \(X_1\) and we prove the following:

**Claim 6:**

\[
\paCor(Z, X_1|\mathcal{X}\setminus X_1, Y_{12}, \ldots, Y_{1n}) = \frac{a_1}{\sqrt{(1 + \sum_{k=2}^{n} \mu_k^2)(\sum_{1\leq i<j\leq n, (i,j)\notin \{(1,2), \ldots, (1,k)\}} b_{ij}^2 + \sigma_i^2) + a_1}}.
\]

**Proof of Claim 6:** By induction, using claim 5 and the recursive formula, one can show that for each interactor \(Y_{ij}\) added in the regression step, its corresponding term in the expression of the \(\paCor\) will cancel out. More precisely, for \(k \in \{1, \ldots, n-1\}\), we can show that:

\[
\paCor(Z, X_1|\mathcal{X}\setminus X_1, Y_{12}, \ldots, Y_{1k}) = \frac{(a_1 + \sum_{j=k+1}^{n} b_{1j}\mu_j)}{\sqrt{(1 + \sum_{j=k+2}^{n} \mu_k^2)(\sum_{1\leq i<j\leq n, (i,j)\notin \{(1,2), \ldots, (1,k)\}} b_{ij}^2 + \sigma_i^2) + (a_1 + \sum_{j=k+1}^{n} b_{1j}\mu_j)^2}}.
\]

This uses the following fact, which is also obtained inductively using the recursive formula:

\[
\paCor(X_1, Y_{1k}|\mathcal{X}\setminus X_1, Y_{12}, \ldots, Y_{1k-1}) = \frac{\mu_k}{\sqrt{1 + \sum_{k=2}^{n} \mu_k^2}}.
\]

**Claim 7:** \(\paCor(Z, X_1|\mathcal{X}\setminus X_1, \mathcal{Y}) = \frac{a_1}{\sqrt{a_1^2 + \sigma_1^2(1 + \sum_{k=2}^{n} \mu_k^2)}}\).
Proof of Claim 7: From claim 6, one obtains that:

\[ \text{paCor}(Z, X_1 | \mathcal{X} \setminus X_1, Y_{12}, ..., Y_{1n}, Y_{23}) = \]

\[ \frac{\text{paCor}(Z, X_1 | \mathcal{X} \setminus X_1, Y_{12}, ..., Y_{1n}) - \text{paCor}(Z, Y_{23} | \mathcal{X} \setminus X_1, Y_{12}, ..., Y_{1n}) \text{paCor}(X_1, Y_{23} | \mathcal{X} \setminus X_1, Y_{12}, ..., Y_{1n})}{\sqrt{1 - \text{paCor}^2(Z, Y_{23} | \mathcal{X} \setminus X_1, Y_{12}, ..., Y_{1n})} \sqrt{1 - \text{paCor}^2(X_1, Y_{23} | \mathcal{X} \setminus X_1, Y_{12}, ..., Y_{1n})}}. \]

From claim 3, and since \( X_1, Y_{23} \) are independent, it follows that

\[ \text{paCor}(X_1, Y_{23} | \mathcal{X} \setminus X_1, Y_{12}, ..., Y_{1n}) = 0. \]

From the proof of claim 4,

\[ \text{paCor}(Z, Y_{23} | \mathcal{X}, Y_{12}, ..., Y_{1n}) = \frac{b_{23}}{\sqrt{\sum_{i<j \leq n} b_{ij}^2 + \sigma_e^2}}. \]

The recursive formula gives:

\[ \text{paCor}(Z, Y_{23} | \mathcal{X} \setminus X_1, Y_{12}, ..., Y_{1n}) = \frac{b_{23}}{\sqrt{\sum_{i<j \leq n} b_{ij}^2 + \sigma_e^2}} \sqrt{1 - p^2(Z, X_1 | \mathcal{X} \setminus X_1, Y_{12}, ..., Y_{1n})}. \]

Hence:

\[ \text{paCor}(Z, X_1 | \mathcal{X} \setminus X_1, Y_{12}, ..., Y_{1n}, Y_{23}) = \frac{\sqrt{(1 + \sum_{i=2}^{n} \mu_k^2) / \left( \sum_{i<j \leq n, (i,j) \neq (2,3)} b_{ij}^2 + \sigma_e^2 \right)}}{\sqrt{\sum_{i<j \leq n} b_{ij}^2 + \sigma_e^2}} \frac{a_1}{a_1} \frac{1}{\sqrt{1 - (1 + \sum_{i=2}^{n} \mu_k^2) / \left( \sum_{i<j \leq n, (i,j) \neq (2,3)} b_{ij}^2 + \sigma_e^2 \right)}} \frac{a_2}{a_2} \frac{1}{\sqrt{1 + \sum_{i=1, k \neq 2}^{n} \mu_k^2}} \frac{b_{12}}{b_{12}^2 + \sigma_e^2}. \]

This shows that for each interactor added in the regression step, its corresponding coefficient in the paCor cancels out and the desired formula is proved.

Proof of the Main Claim: From claims 1 – 7, we have obtained the following explicit formulas:

\[ \text{paCor}(Z, X_1 | \mathcal{X} \setminus X_1, \mathcal{Y}) = \frac{a_1}{\sqrt{a_1^2 + \sigma_e^2(1 + \sum_{k=2}^{n} \mu_k^2)}}. \]

\[ \text{paCor}(Z, X_2 | \mathcal{X} \setminus X_2, \mathcal{Y}) = \frac{a_2}{\sqrt{a_2^2 + \sigma_e^2(1 + \sum_{k=1, k \neq 2}^{n} \mu_k^2)}}. \]

\[ \text{paCor}(Z, Y_{12} | \mathcal{X}, \mathcal{Y} \setminus Y_{12}) = \frac{b_{12}}{\sqrt{b_{12}^2 + \sigma_e^2}}. \]

Then

\[ |\text{paCor}(Z, Y_{12} | \mathcal{X}, \mathcal{Y} \setminus Y_{12})| \geq \max(\|\text{paCor}(Z, X_1 | \mathcal{X} \setminus X_1, \mathcal{Y})\|; \|\text{paCor}(Z, X_2 | \mathcal{X} \setminus X_2, \mathcal{Y})\|) \text{ if and only if } |b_{12}| \geq \max\left( \frac{|a_1|}{\sqrt{1 + \sum_{k=2}^{n} \mu_k^2}}; \frac{|a_2|}{\sqrt{1 + \sum_{k=1, k \neq 2}^{n} \mu_k^2}} \right). \]
Acknowledgements: This work was supported in part by R01GM076274 to CK. The authors wish to thank Michael Newton for comments that greatly broadened the presentation of this work.

REFERENCES


