ORIGINAL ARTICLE
Journal Section

Corrected score methods for estimating Bayesian networks with error-prone nodes

Xianzheng Huang[1] | **Hongmei Zhang**[2]

[1]Department of Statistics, University of South Carolina, Columbia, SC, 29208, U.S.A.

[2]Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN, 38152, U.S.A

Correspondence
Xianzheng Huang, Department of Statistics, University of South Carolina, Columbia, SC, 29208, U.S.A.
Email: huang@stat.sc.edu

Funding information

Motivated by inferring cellular signaling networks using noisy flow cytometry data, we develop procedures to draw inference for Bayesian networks based on error-prone data. Two methods for inferring causal relationships between nodes in a network are proposed based on penalized estimation methods that account for measurement error and encourage sparsity. We discuss consistency of the proposed network estimators and develop an approach for selecting the tuning parameter in the penalized estimation methods. Empirical studies are carried out to compare the proposed methods with a naive method that ignores measurement error. Finally, we apply these methods to infer signaling networks using single cell flow cytometry data.

KEYWORDS
false discovery rate, Frobenius norm, information criterion, specificity, topological sorting

1 | INTRODUCTION

1.1 | Motivations

The study of cellular signaling networks has been a major research area in biology for several decades. By analyzing how multiple cell signaling pathways affect each other within a network, scientists gain valuable insights on normal cellular responses in a biological system, and their potential disregulation in disease [1, 2, 3]. Statistical models that mathematically conceptualize these signaling networks have been developed [4, 5]. These models have advanced experimental cell biology, and influenced the way biologists view, monitor, and study signaling networks by perturbing them in designed experiments [6, 7].

Among these models, Bayesian networks [8] have been widely adopted as an attractive model for characterizing complex cell signaling cascades. With recent advances in biochemistry, molecular biology, and cell physiology, rich data information become available at the cell level from high throughput technologies. For example, flow cytometry makes measuring physical and chemical characteristics of cells possible, and has become an important tool in a broad range of biological and clinical research. This technology produces abundant data that can be used to infer cellular signalling networks [9, 10, 11, 12, 13]. However, measurement errors in flow cytometry data inevitably arise from imperfect measurements, photon-counting statistics, and data storage methods [14, 15, 16, 17]. Figure 1, borrowing from Figure 1 in Galbusera et al. [17], illustrates data (such as the height, area, and width of a pulse) reported by a cytometer. To collect such data, one streams cells past a laser light source, and detects them via fluorescence picked up by detectors. The fluorescence signal often contains autofluorescence resulting from the laser exciting cellular components other than the green fluorescent protein that one intends to excite. Automated compensation tools have been developed to correct for these nuisance sources of fluorenscence, and for spillove resulting from using fluorescent dyes measurable in more than one detector. Yet the autocompensation procedure depends on flow cytometrists' subjective gate placement. Error arising in the compensated data are discussed in greater details in Roederer [14]. Besides this and earlier references, the following blog gives a more detailed take on "bad flow cytometry data" and spectral compensation in the context of more recent models of flow cytometers, `https://voices.uchicago.edu/ucflow/author/lkjohnston`. These practical concerns regarding the quality of flow cytometry data even with the latest technology motivate our study presented in this article, where we develop methods for inferring Bayesian networks representing cellular signaling networks using error-prone flow cytometry data.

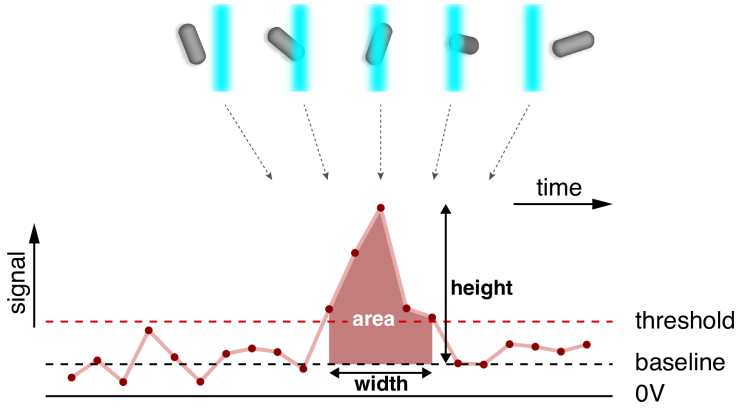

FIGURE 1 An illustration of the signals reported by the cytometer given in Figure 1 in Galbusera et al. [17].

The proposed methods can be used to infer Bayesian networks arising in other applications. Examples of these applications include constructing social networks based on survey data subject to imperfect respondent recall [18], studying connectivity and association between different regions of one's brain in the default mode network using preprocessed noisy brain image data [19], and modeling gene regulatory pathways using gene expression data that are imprecise due to experimental errors [20], stray background signal irrelevant to mRNA transcripts [21], or data normalization [22]. With this wide range of applications in mind, we provide a literature overview next on networks and network inference in a general context.

1.2 | Literature Review

Networks, or graphs, have been a topic of great interest that started mostly in the artificial intelligence community [8, 23, 24]. Later their application became more widespread, motivating statistical research on graphical models used in biology, genetics, social science, and physics [25, 26]. A network consists of a set of nodes, also referred to as vertices or variables, and a set of edges connecting nodes. Graphs with undirected edges are called undirected graphs. In an undirected graph, a set of nodes connecting to a particular node form a neighborhood of this node. Given its neighborhood, this node is independent of nodes outside of the neighborhood. This type of graphs is useful for characterizing correlations between nodes. When causal relationships are of interest, directed edges are used, giving rise to the so-called directed acyclic graphs (DAG). Pairing such a graph with a joint probability distribution of all nodes produces a Bayesian network. When there is an edge pointing from one node to another node, these two nodes are referred to as a parent node (of the latter) and a child node (of the former), respectively. Given its parents, a node is independent of its non-descendant nodes. This is more formally known as the local Markov property of DAG, which in turn suggests that a Bayesian network encodes the joint distribution of the set of nodes in the graph. Provided with this encoding, not only can one uncover the correlation structure among nodes, but one can also reveal if a correlation between two nodes is due to a direct causal relationship between them or an indirect dependence mediated by other nodes. The latter piece of information is especially of interest in biology and genetics [27]. Because of this, some researchers refer to Bayesian networks as causal networks to signify causality as their research focal point, such as in research on cellular signaling networks.

There is a large collection of works on inferring Bayesian networks. Many existing works follow the theme of search-and-score [28, 29, 30, 31, 32, 33, 34, 35]. Following this theme, one formulates a scoring criterion, and searches for a directed graph, or an equivalent class of directed graphs [36], that optimizes the score. The score can be constructed based on a likelihood function of observed data in the frequentist framework [11, 37]; it can also originate from a posterior distribution of a graph in the Bayesian framework [29]. Scores formulated borrowing these two schools of statistics have also been used, such as the Bayesian Dirichlet equivalent uniform score defined as the log likelihood of the observed data given suitably chosen Dirichlet priors over the parameters of a network structure [35]. When the number of nodes is large, scores designed to penalize complexity of a graph are often employed [38, 39, 40]. Another well explored theme for inferring Bayesian networks leads to the constraint-based approaches that involve testing conditional independence among nodes [41, 42]. To lessen the computational burden in the presence of many nodes, Tsamardinos et al. [43] proposed the max-min hill-climbing algorithm that combines ideas from search-and-score, constraint-based approaches, and local learning. Friedman and Koller [44] used a Markov chain Monte Carlo (MCMC) method over the space of node orders, which is smaller and more regular than the space of graph structures. Eaton and Murphy [45] suggested to apply dynamic programming algorithm on the space of node orders, then used the resultant proposal distribution for MCMC methods in the space of DAGs. Also considering the order space, Ellis and Wong [46] developed a fast MCMC algorithm based on data that include interventional data and observational data. Ye et al. [47] proposed to minimize a regularized Cholesky score over the space of topological orderings. When compared with several competing methods applied to both observational and interventional data, their method achieved improved performance in network structure learning. Interventional data arise from intervention experiments, such as flow cytometry experiments considered in our study. In such an experiment, one fixes the values of some node(s), which in effect destroys the causal dependencies of the intervened node(s). Inclusion of interventional data greatly improves the identifiability of a Bayesian network, as Hauser and Bühlmann [48] and Peters et al. [49] explained in great detail.

All aforementioned existing works rely on observed data as precise measures of nodes. But, as seen in the moti-

vating examples, measures of nodes can be imprecise. For flow cytometry experiments, Galbusera et al. [17] showed that flow cytometry measurements contain a significant amount of shot-noise that can be easily mistaken for true biological variability. Although measurement error problems have been long investigated in many regression settings [50, 51, 52], there is very limited research in the context of inferring Bayesian networks. One exception is the work by Luo and Zhao [12], who used Bayesian hierarchical modeling to incorporate measurement error and random error that represent intrinsic noise in flow cytometry data when inferring signaling pathways. In this article, we tackle this problem from the frequentist point of view. To the best of our knowledge, this is the first frequentist work addressing this problem.

The data structure considered in our study and mathematical formulations of the data generating mechanism are described in Section 2. We then outline the proposed penalized estimation methods in Section 3, which includes detailed algorithms for implementing the proposed methods. To choose the tuning parameter in the penalized estimation, we construct a tuning parameter selector in Section 4. In Section 5, simulation studies are reported, where we compare finite sample performance of the proposed methods and a naive method that ignores measurement error, and carry out sensitivity analysis under incorrect assumptions on measurement errors. We also apply these methods to a flow cytometry data set to infer a signaling network of immune system cells. In Section 6, we summarize the contribution of our study and discuss follow-up research.

2 | DATA AND MODEL

Denote by $\mathbf{X}$ the $N \times p$ (unobserved) data matrix as error-free measures of p nodes in a network, including interventional data and observational data from N experimental units. Refer to node j as X_j, denote by n_j and n_{-j} the number of interventional data points and the number observational data points associated with X_j, respectively, so that $n_j + n_{-j} = N$, and by O_j the set of row indices corresponding to the observational data for X_j in $\mathbf{X}$, for $j = 1, \ldots, p$. If an interventional study involves p experimental conditions under each of which a different node is intervened, then $\sum_{i=1}^{p} n_j = N$. In general, $\sum_{i=1}^{p} n_j$ can be below or above N, e.g., when some nodes are never intervened or are intervened in more than one condition. The observed data matrix of the same dimension, $\mathbf{W}$, is an error-contaminated surrogate of $\mathbf{X}$.

Taking the data structure into consideration, we assume that the causal relationships of the p nodes are specified by

$$\mathbf{X}[O_j, j] = \mathbf{X}[O_j, -j]\mathbf{B}_j + \boldsymbol{\epsilon}[O_j, j], \text{ for } j = 1, \ldots, p, \tag{1}$$

where $\mathbf{X}[O_j, j]$ is the $n_{-j} \times 1$ vector taken from the jth column of $\mathbf{X}$ containing only entries in the rows indicated by O_j, $\mathbf{X}[O_j, -j]$ is an $n_{-j} \times (p-1)$ submatrix of $\mathbf{X}$ storing data in the rows indicated by O_j and all columns except the jth column, $\boldsymbol{\epsilon}$ is the $N \times p$ matrix of model error representing intrinsic noise due to unmodelling variation, $\boldsymbol{\epsilon}[O_j, j]$ consists of n_{-j} independent and identically distributed (i.i.d.) mean-zero random errors, $\mathbf{B} = [\beta_{ij}]_{i,j=1,\ldots,p}$ is the $p \times p$ matrix of regression coefficients with zero diagonal entries, and $\mathbf{B}_j = \mathbf{B}[-j, j]$. The regression model representation of a Bayesian network in (1) is the same as that formulated in Fu and Zhou [13]. It is assumed that $\boldsymbol{b} = (\mathbf{B}_1^{\mathsf{T}}, \ldots, \mathbf{B}_p^{\mathsf{T}})^{\mathsf{T}}$ is a vector of natural parameters in the sense that, given sufficient interventional data associated with each node, $\boldsymbol{b}$ is identifiable [13]. For X_j, the nodes on the right-hand-side of (1) associated with nonzero entries in $\mathbf{B}_j$ are parents of X_j. Having $\mathbf{B}_j = \mathbf{0}$ means that X_j has no parent, and is referred to as a root node. Having the jth row, $\mathbf{B}[j,]$, as a zero vector implies that X_j has no child, and is referred to as a leaf node. Assume that $\mathbf{W}$ results from contaminating

X with additive mean-zero normal measurement error independent of **X**, that is,

$$\mathbf{W} = \mathbf{X} + \mathbf{U}, \tag{2}$$

where **U** is the $N \times p$ matrix of nondifferential measurement error [50, Section 2.5]. It is further assumed in this study that, for each node X_j, the measurement error associated with the interventional data of X_j and the measurement error associated with the observational data of X_j follow the same distribution. This implies that $\{\mathbf{U}[\ell, \ ]\}_{\ell=1}^{N}$ are i.i.d. random vectors from $N_p(\mathbf{0}, \mathbf{\Sigma}_u)$, where $\mathbf{\Sigma}_u$ is the $p \times p$ variance-covariance matrix of the measurement error associated with nodes $(X_1, \ldots, X_p)$. This assumption is practically reasonable when the source of measurement error remains the same throughout an experiment, for instance, by using the same model of flow cytometers in all experimental conditions.

According to (1) and (2), the Bayesian network with error-prone nodes consists of p hierarchical measurement error models, with the jth hierarchical model consisting of two submodels,

$$\mathbf{W}[O_j, j] = \mathbf{X}[O_j, -j]\mathbf{B}_j + \boldsymbol{\epsilon}[O_j, j] + \mathbf{U}[O_j, j], \tag{3}$$

$$\mathbf{W}[O_j, -j] = \mathbf{X}[O_j, -j] + \mathbf{U}[O_j, -j], \tag{4}$$

where the first submodel is for the error-contaminated node j regressing on the remaining $p - 1$ error-free nodes, and the second submodel relates the observed covariates with the true covariates in the jth regression model, for $j = 1, \ldots, p$. Given the set of p measurement error models, making inference for an underlying Bayeisan network that relates the p nodes mainly involves inferring **B** using **W**. The variance-covariance associated with $\boldsymbol{\epsilon}[O_j, j]$ does not need to be estimated for the proposed methods. Estimating $\mathbf{\Sigma}_u$ requires either external validation data or replicate measures of the same set of error-free measures of nodes [50]. Such estimation has been a routine practice in the measurement error literature, where researchers mostly report little impact on the final inference on regression parameters. In order to focus on inference on **B**, we assume $\mathbf{\Sigma}_u$ known in the methodology development, and investigate consequences of estimating or misspecifying $\mathbf{\Sigma}_u$ in empirical studies in Section 5.

3 | ESTIMATION OF B

3.1 | Penalized Objective Functions

When **X** is observed, Fu and Zhou [13] proposed to estimate **B** via maximizing a penalized log-likelihood function corresponding to the graphical model in (1). More recently, Li et al. [37] developed likelihood ratio tests for connectivity and directionality of a DAG that involve maximizing constrained and penalized log-likelihood functions built upon regression models in (1). In the presence of measurement error, a naive likelihood-based approach for estimating **B** is to ignore measurement error and use **W** in place of **X** in the penalized negative log-likelihood function as follows,

$$R_{\mathrm{nv}}(\mathbf{B}) = \sum_{j=1}^{p} \left\{ V_{j,\mathrm{nv}} + \sum_{i=1}^{p} P_\lambda(|\beta_{ij}|) \right\}, \tag{5}$$

where, for $j = 1, \ldots, p$,

$$V_{j,\mathrm{nv}} = \frac{n_{-j}}{2} \log \left\{ \sum_{\ell \in O_j} \left(\mathbf{W}[\ell, j] - \mathbf{W}[\ell, -j]\mathbf{B}_j \right)^2 \right\}, \tag{6}$$

and $P_\lambda(\cdot)$ is a penalty function. One may choose a penalty according to the LASSO [53], the adaptive LASSO (ALASSO) [54], or the SCAD penalty [55]. Both ALASSO and SCAD have been shown to enjoy the appealing oracle properties in variable selections. In this study we adopt SCAD in (5), defined as

$$P_\lambda(t) = \lambda t I(t \in [0, \lambda)) + \frac{(a^2 - 1)\lambda^2 - (t - a\lambda)^2}{2(a - 1)} I(t \in [\lambda, a\lambda)) + \frac{(a + 1)\lambda^2}{2} I(t \geq a\lambda),$$

where λ is a tuning parameter, $I(\cdot)$ is the indicator function, and $a = 3.7$. Besides avoiding the adaptive weights required in ALASSO, our choice of the SCAD penalty is also motivated by findings in Aragam and Zhou [56]. There, the authors showed that a concave penalty, such as SCAD, offers improved performance in Bayesian network structure learning when comparing with an L_1-based penalty like LASSO. Minimizing (5) with respect to $\mathbf{B}$ yields an estimator of it, denoted by $\hat{\mathbf{B}}_{\mathrm{nv}}$.

To account for measurement error in node data, we construct a penalized objective function based on the corrected score function [57]. Assuming normal model error and measurement error, the corrected score function associated with the jth measurement error model is given by

$$\begin{aligned} \boldsymbol{\Psi}_j(\mathbf{B}_j) &= \sum_{\ell \in O_j} \boldsymbol{\Psi}_{j\ell}(\mathbf{B}_j) \\ &= \sum_{\ell \in O_j} \left\{ (\mathbf{W}[\ell, j] - \mathbf{W}[\ell, -j]\mathbf{B}_j)\mathbf{W}[\ell, -j]^t + \boldsymbol{\Sigma}_u[-j, -j]\mathbf{B}_j \right\}. \end{aligned} \tag{7}$$

When $\boldsymbol{\Sigma}_u = \mathbf{0}$, the summand in (7) reduces to the score used in the least squares method for estimating the regression coefficients in the jth regression model, for $j = 1, \ldots, p$. In the presence of measurement error, one can show that $E\{\boldsymbol{\Psi}_{j\ell}(\mathbf{B}_j^*)\} = \mathbf{0}$ [50, Section A.6], where $\mathbf{B}_j^*$ is the truth of $\mathbf{B}_j$, for $\ell \in O_j$ and $j = 1, \ldots, p$. In other words, the corrected score, $\boldsymbol{\Psi}_{j\ell}(\mathbf{B}_j)$, is an unbiased score that corrects the naive least squares score for measurement error.

For each $j \in \{1, \ldots, p\}$, we follow the construction of quadratic inference functions [58] and propose the penalized score-based objective function given by

$$R(\mathbf{B}) = \sum_{j=1}^{p} \left\{ V_j + \sum_{i=1}^{p} P_\lambda(|\beta_{ij}|) \right\}, \tag{8}$$

where

$$V_j = \left\{ \frac{1}{n_{-j}} \sum_{\ell \in O_j} \boldsymbol{\Psi}_{j\ell}(\mathbf{B}_j) \right\}^t \{\mathbf{H}_j(\mathbf{B}_j)\}^{-1} \left\{ \frac{1}{n_{-j}} \sum_{\ell \in O_j} \boldsymbol{\Psi}_{j\ell}(\mathbf{B}_j) \right\}, \tag{9}$$

in which $\mathbf{H}_j(\mathbf{B}_j) = n_{-j}^{-1} \sum_{\ell \in O_j} \boldsymbol{\Psi}_{j\ell}(\mathbf{B}_j) \boldsymbol{\Psi}_{j\ell}(\mathbf{B}_j)^t$ is a consistent estimator for the variance-covariance matrix of the corrected score, which is "sandwiched" between the scores to achieve optimal efficiency in the score-based inference [59]. A non-naive estimator of $\mathbf{B}$, denoted by $\hat{\mathbf{B}}$, is a minimizer of $R(\mathbf{B})$.

In Appendix A of the Supplementary Materials, we establish the consistency of the estimator as a minimizer of (8) with a fixed p under regularity conditions. Denote by $\mathbf{B}^*$ the true value of $\mathbf{B}$, and define $\boldsymbol{b}^* = (\mathbf{B}_1^{*t}, \ldots, \mathbf{B}_p^{*t})^t$. Write the tuning parameter λ in (8) as λ_n to signify its dependence on $n = \min_{1\le j\le p} n_{-j}$ in the discussion of asymptotics. The consistency of $\hat{\mathbf{B}}$ is stated in the following theorem.

Theorem 3.1 Under assumptions (A1)–(A5) in Appendix A, as $n \to \infty$, if $\sqrt{n}\lambda_n = o_p(1)$, then there exists a local minimizer of $R(\mathbf{B})$ defined in (8), denoted by $\hat{\mathbf{B}}$, such that $\|\hat{\boldsymbol{b}} - \boldsymbol{b}^*\| = O_p(n^{-1/2})$, where $\hat{\boldsymbol{b}} = (\hat{\mathbf{B}}_1^t, \ldots, \hat{\mathbf{B}}_p^t)^t$.

3.2 | Algorithms for Estimating B

To find an optimizer of the penalized log-likelihood, Fu and Zhou [13] developed a pairwise coordinate descent (PCD) algorithm to iteratively update each of the $p(p-1)/2$ pairs, (β_{ij}, β_{ji}), for $i \neq j = 1, \ldots, p$, with all other entries of $\mathbf{B}$ fixed at their values from the preceding iteration. The algorithm is designed to avoid estimates for $\beta_{ij}$and β_{ji} to be nonzero simultaneously, since β_{ij} and β_{ji} both being nonzero is a violation of acyclicity. But PCD does not check for other forms of acyclicity violation. To thoroughly check for cycles in an estimated regression coefficients matrix, we implement Kahn's topological sorting algorithm [60] along with PCD.

Given a directed graph structure G, a topological sorting algorithm is an iterative procedure that yields a sorted sequence of nodes in which a child node always comes after its parent nodes, and thus specifies a topological ordering of these nodes compatible with G. A topological sorting algorithm can be used to detect cycles because a topological ordering of nodes does not exist whenever there exists a cycle in the graph [61]. In particular, Kahn's sorting algorithm is developed based on the fact that a DAG must have at least one root node; moreover, removing root nodes and their out-going edges from a DAG always yields a subgraph that is still a DAG. Hence, an early termination of the sorting algorithm will only occur if a subgraph at that step has no root node, which indicates existence of at least one cycle in the (sub)graph. When this occurs, we will strategically remove edges until root nodes emerge so that the sorting algorithm can resume. Figure 2 illustrates the application of Kahn's sorting algorithm for the purpose of cycle detection and elimination for an initial graph structure G. The output of the depicted algorithm is an order compatible with the resultant acyclic graph specified by a queue of p nodes, denoted by $\mathscr{T}$. At the beginning of the algorithm, we set $\mathscr{T}$ as an empty queue, and accumulate in $\mathscr{T}$ the root nodes of the (sub)graphs created during the iterative procedure.

In Figure 2, the weakest edge in G mentioned in the middle gray-shaded box is the edge associated with an estimated regression coefficient that indicates the weakest association among all non-zero estimated regression coefficients. We use p-values of the estimated regression coefficients to identify the weakest edge to be removed until the sorting algorithm resumes due to newly emerging root nodes. By the time the queue $\mathscr{T}$ collects all p nodes, we obtain a final regression coefficient matrix estimate by placing zeros in the entries corresponding to the removed weak edges.

A complete algorithm for finding a minimizer of the penalized score-based objective function $R(\mathbf{B})$ in (8) that specifies a DAG is related next, which uses the PCD algorithm in conjunction with Kahn's sorting algorithm.

Step 1: Obtain an initial estimate of $\mathbf{B}$ by solving p unpenalized corrected score estimating equations one at a time. Denote by $\hat{\mathbf{B}}^{(0)}$ the resultant initial estimate of $\mathbf{B}$. Set the iteration index $t = 0$.

Step 2: For $i, j \in \{1, \ldots, p\}$ and $i \neq j$, define $\tilde{\beta}_{ij} = \hat{\mathbf{B}}^{(t)}[i, j]$ and $\tilde{\beta}_{ji} = \hat{\mathbf{B}}^{(t)}[j, i]$. For each pair of nodes i and j, update $(\tilde{\beta}_{ij}, \tilde{\beta}_{ji})$ to $(\tilde{\beta}_{ij}^*, \tilde{\beta}_{ji}^*)$ by minimizing the penalized score-based objective function following the algorithm elaborated in Appendix B of the Supplementary Materials. Set $\hat{\mathbf{B}}^{(t+1)} = [\tilde{\beta}_{ij}^*]_{i,j=1,\ldots,p}$. Denote by $\tilde{G}$ the graph structure specified by $\hat{\mathbf{B}}^{(t+1)}$, which may have cycles.

Step 3: For $j = 1, \ldots, p$, compute unpenalized corrected score estimates for regression coefficients associated with

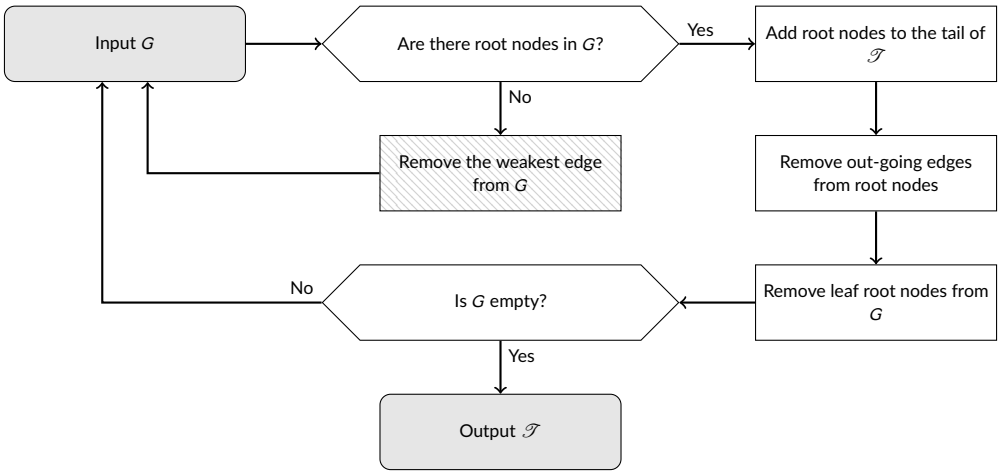

FIGURE 2 Kahn's topological sorting algorithm for eliminating cycles in G and finding a topological ordering, $\mathscr{T}$, compatible with the resultant acyclic graph.

the parents of X_j indicated by $\tilde{G}$. Obtain estimated standard errors of these unpenalized regression coefficients estimates via sandwich variance estimation for M-estimators. Produce p-values based on the unpenalized estimate of β_{ij} and its estimated standard error for testing $H_0 : \beta_{ij} = 0$ versus $H_1 : \beta_{ij} \neq 0$, if X_i is a parent of X_j according to $\tilde{G}$.

Step 4: Implement Kahn's sorting algorithm to eliminate cycles in $\tilde{G}$ by setting some (initially nonzero in Step 3) coefficients in $\hat{\mathbf{B}}^{(t+1)}$ to zero that have the largest p-values, unless $\tilde{G}$ from Step 3 is a DAG.

Step 5: If $|\hat{\mathbf{B}}^{(t+1)} - \hat{\mathbf{B}}^{(t)}|_\infty$ is larger than a small pre-sepecified threshold, set $t = t+1$, and return to Step 2. Otherwise, output $\hat{\mathbf{B}}^{(t+1)}$ as a minimizer of $R(\mathbf{B})$ that specifies a DAG. Here, for a matrix $\mathbf{A}$, $|\mathbf{A}|_\infty$ denotes the largest entry of $\mathbf{A}$ in absolute value.

The threhold we set in Step 5 is 10^{-4} in our simulation study, and we observe negligible change in the output of the algorithm when an even smaller threshold is used. One can follow a similar algorithm described above to find the optimizer of the naive penalized log-likelihood function $R_{\mathrm{nv}}(\mathbf{B})$ in (5) that relates to a DAG. This is elaborated in Appendix C of the Supplementary Materials, where formulas for updating each pair of regression coefficients in Step 2 are provided. The algorithm implemented in Fu and Zhou [13] to optimize their penalized log-likelihood function using error-free data does not include Steps 3 and 4 above and thus does not guarantee to return a DAG in the end. Li et al. [37] applied the alternating direction method of multiplier (ADMM) [62] to update parameters in their penalized objective function, followed by depth-first search [61] for cycle detection, and then deleted weak edges based on the absolute values of estimated regression coefficients to ensure acyclicity of a graph at each iteration.

When implementing the PCD algorithm, one considers one pair of regression models at a time within each iteration: the regression model with X_j as the response and the one with X_i as the response. For each pair of models, one focuses on inferring one regression coefficient in each model in that iteration. In particular, one chooses between "X_i is an influential covariate in the jth regression model" and "X_j is an influential covariate in the ith regression model," given all other covariates chosen from the previous iteration for that model. Alternatively, instead of updating $\hat{\mathbf{B}}^{(t)}$ one pair of entries at a time, one may update one column of $\hat{\mathbf{B}}^{(t)}$ at a time by selecting important covariates for the jth regression model, for $j = 1, \ldots, p$. This leads to another approach for estimating $\mathbf{B}$ that follows a similar algorithm

but with the following step replacing Step 2 above.

Step 2^***:** For $j = 1, \ldots, p$, use $\hat{\mathbf{B}}_j^{(t)}$ as the starting value to solve the following penalized score estimating equation,

$$n_{-j}^{-1} \sum_{\ell \in O_j} \boldsymbol{\Psi}_{j\ell}(\mathbf{B}_j) - \tilde{P}_\lambda(\mathbf{B}_j) = \mathbf{0}, \tag{10}$$

where $\tilde{P}_\lambda(\mathbf{B}_j)$ is a $(p-1) \times 1$ vector with entries given by, for $k \neq j$,

$$\frac{\partial}{\partial \beta_{kj}} P_\lambda(|\beta_{kj}|) = \lambda \left\{ I(|\beta_{kj}| \leq \lambda) + \frac{(a\lambda - |\beta_{kj}|)_+}{(a-1)\lambda} \right\} \operatorname{sign}(\beta_{kj}).$$

Let the resultant p sets of solutions be the p columns in $\hat{\mathbf{B}}^{(t+1)}$, though not including the zero diagonal element in each column. Denote by $\tilde{G}$ the graph structure induced by $\hat{\mathbf{B}}^{(t+1)}$.

We use Newton-Raphson algorithm to solve (10), where the derivative of $\tilde{P}_\lambda(\mathbf{B}_j)$ is approximated by a $(p-1) \times (p-1)$ diagonal matrix whose diagonal entries are given by $I(\beta_{kj} \neq 0)|(\partial/\partial\beta_{kj})P_\lambda(|\beta_{kj}|)|/|\beta_{kj}|$, for $k \neq j$. This is also the local quadratic approximation used in Fan and Li [55]. We refer to this algorithm as the node-wise parent selection (NPS) algorithm to distinguish it from the previous algorithm that involves PCD. Unlike Step 2 in the PCD algorithm, Step 2^* here ignores the acyclicity constraint. Steps 3 and 4 are where we impose this constraint in the NPS algorithm.

For each node, the NPS algorithm in Step 2^* is precisely the algorithm proposed by Huang and Zhang [63] for variable selection in one linear regression model with error-prone covariates. Ma and Li [64] developed a similar score-based variable selection strategy applicable to a larger class of measurement error models. Both groups of authors established consistency of the score-based variable selection method when one regression model is considered. Without considering the correlation between p regression models that the Bayesian network decomposes into, we expect that this alternative algorithm can yield a sensible estimate for $\mathbf{B}$ that ignores the acyclicity constraint, and the cycle detection and elimination in Step 4 allows one to impose this constraint on the output of the NPS algorithm. Putting the penalty term aside, solving the p sets of penalized score estimation equations in (10) is intrinsically related to minimizing the penalized score-based objective function in (8) since they both originate from the corrected score.

4 | TUNING PARAMETER SELECTION

We are now in the position to discuss choices of the tuning parameter λ in the penalized score-based objective function in (8) and the penalized score estimating equation in (10). In principle, it is desirable to use a consistent information criterion to choose λ. Assume that the class of candidate models includes the true model which the observed data come from. In the context of variable selection in a regression model, within the class of all candidate models, a correct model includes all truly influential predictors in the true model and may also include non-influential predictors; the rest are incorrect models that are referred to as underfitted models. In other words, the true model is the most parsimonious correct model, and a correct model that is not the true model is an overfitted model. A consistent information criterion refers to a criterion that approaches (in probability) to its optimal value as the sample size tends to infinity when evaluated at the true model.

To infer a Bayesian network consisting of error-prone nodes, we propose the score-based information criterion

evaluated at a graph G given by

$$\mathrm{SIC}(G) = \sum_{j=1}^{p} \left(\hat{V}_j + e_j \frac{\log n_{-j}}{n_{-j}} \right), \tag{11}$$

where e_j is the number of parents of X_j according to G, and $\hat{V}_j$ is equal to V_j evaluated at the unpenalized corrected score estimate of $\mathbf{B}_j$ given the structure of G. In the context of linear regression with error-prone covariates, Huang and Zhang [63] developed two score-based information criteria very much in the same spirit as the summand in (11) to facilitate variable selection in one regression model. The proposed information criterion in (11) is essentially the sum of p score-based information criteria associated with p regression models as the decomposition of a Bayesian network. To establish its consistency as a model criterion, it is instructive to relate arguments for model selection where one regression model is concerned to arguments for graph selection, where a graph can be decomposed into p regression models.

Denote by $\mathbf{E}_G$ the set of directed edges in G, and by $|\mathbf{E}_G|$ the size of this set. Suppose there exists a true Bayesian network that dictates the data generating process, with graph structure G_0, in the class of networks under consideration. Parallel with notions in variable selection in the regression setting, let G_- and G_+ denote generically an underfitted graph and an overfitted graph, respectively, where G_- satisfies $\mathbf{E}_{G_0} \not\subset \mathbf{E}_{G_-}$, and G_+ satisfies $\mathbf{E}_{G_0} \subset \mathbf{E}_{G_+}$. Then G_0 and G_+ are correct graphs, with the former more parsimonious than the latter, that is, $|\mathbf{E}_{G_0}| < |\mathbf{E}_{G_+}|$. In contrast, G_- is an incorrect graph, and one does not necessarily have $|\mathbf{E}_{G_-}| < |\mathbf{E}_{G_0}|$. To establish the consistency of $\mathrm{SIC}(G)$, it suffices to show that

$$\begin{aligned} \mathrm{SIC}(G_-) - \mathrm{SIC}(G_0) \quad &> 0 \quad \text{with probability approaching one, as } n \to \infty \text{ and,} \\ \mathrm{SIC}(G_+) - \mathrm{SIC}(G_0) \quad &\to 0^+ \quad \text{in probability as } n \to \infty. \end{aligned}$$

These assertions are proved in Appendix D of the Supplementary Materials, where p is allowed to diverge as $n \to \infty$.

5 | EMPIRICAL EVIDENCE

5.1 | Competing Methods

In this section, we implement the proposed score-based methods and the naive likelihood-based method using simulated network data to assess their finite sample performance. For the score-based methods, we use the SIC tuning parameter selector to choose λ. For the naive method, we adopt the tuning parameter selection method employed in Fu and Zhou [13] based on the relative change in prediction error.

Denote by e_λ the number of edges of an estimated graph when the tuning parameter is set to λ, and by $\hat{\mathbf{B}}_{\mathrm{nv}}^{(\lambda)}$ the corresponding naive estimate of $\mathbf{B}$. Define the prediction error by $\mathrm{PE}_\lambda = \sum_{j=1}^{p} \sum_{\ell \in O_j} (\mathbf{W}[\ell, j] - \hat{\mathbf{W}}^{(\lambda)}[\ell, j])^2$, where $\hat{\mathbf{W}}^{(\lambda)}[\ell, j] = \mathbf{W}[\ell, -j]\hat{\mathbf{B}}_{\mathrm{nv}\,j}^{(\lambda)}$. Suppose one considers m candidate values for λ, $\lambda_1 > \lambda_2 > \ldots > \lambda_m$. For each $k = 2, \ldots, m$, one computes the relative change in prediction error defined by $\mathrm{RCP}_{k-1,k} = (\mathrm{PE}_{\lambda_{k-1}} - \mathrm{PE}_{\lambda_k})/(e_{\lambda_k} - e_{\lambda_{k-1}})$, if $e_{\lambda_k} - e_{\lambda_{k-1}} > 0$, and $\mathrm{RCP}_{k-1,k} = 0$ otherwise. Then one chooses λ_K as the tuning parameter value, where $K = \max\{k : \mathrm{RCP}_{k-1,k} \geq \alpha \max(\mathrm{RCP}_{1,2}, \ldots, \mathrm{RCP}_{m-1,m}),\ k = 2, \ldots, m\}$, in which α is a threshold parameter set to 0.1 as recommended by Fu and Zhou [13]. The quantity defined as $\mathrm{RCP}_{k-1,k}$ quantifies the gain in prediction accuracy at the price of increasing graph complexity (as e_λ increases) when one drops λ from λ_{k-1} to λ_k. The use of the threshold α is to further guard against overly dense graphs. The constructions of RCP and K together aim to

balance graph complexity and prediction accuracy.

In summary, there are three methods implemented in the simulation study: the naive likelihood-based method using the PCD algorithm with λ chosen by RCP, the score-based method using the PCD algorithm with λ chosen by SIC, and the score-based method using the NPS algorithm with λ chosen by SIC.

5.2 | Simulation Settings

The simulation experiment involves two factors: the number of nodes p and the variance-covariance matrix of the measurement error $\mathbf{\Sigma}_u$. There are two levels for p, 10 and 20. Given p, the total number of edges of a true graph is $3p$, and each node has at most four parents. Once such a graph is created randomly, we set the entries in $\mathbf{B}$ associated with the first half of edges to 0.5, and entries associated with the second half of edges to 1. Then we generate $n_j = 5$ interventional data points from $N(0,1)$, for $j = 1,\ldots,p$. When generating normal measurement errors, we first set $\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{I}_p$, where σ_u^2 varies across five levels so that the reliability ratio, $\tau = \mathrm{Var}(X_j)/\{\mathrm{Var}(X_j) + \sigma_u^2\}$, associated with each X_j ranges from 0.8 to 1 at increments of 0.05, for $j = 1,\ldots,p$. In a different setting we let $\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{V}_p$, where σ_u^2 takes the five aforementioned levels, and $\mathbf{V}_p$ is a $p \times p$ matrix with entries given by $\mathbf{V}_p[j,j'] = 0.5^{|j-j'|}$, for $j,j' = 1,\ldots,p$. For each simulation setting, we randomly generate ten graphs, from each of which an $N \times p$ data matrix $\mathbf{W}$ is generated according to (3) and (4) with $\{\boldsymbol{\epsilon}[\ell,j], \ell = 1,\ldots,N\}_{j=1}^{p}$ being independent realizations from $N(0,1)$.

Given a true graph G, the following five metrics are used to assess the quality of an estimated graph $\hat{G}$:

- the true positive rate, TPR = $|\mathbf{E}_{\hat{G}} \cap \mathbf{E}_G|/(3p)$;
- the false discovery rate, FDR = $(\mathrm{R} + |\mathbf{E}_{\hat{G}} \cap \mathbf{E}_G^c|)/|\mathbf{E}_{\hat{G}}|$, where R denotes the number of edges in G that show up in $\hat{G}$ in the reversed direction;
- the specificity defined as $|\mathbf{E}_{\hat{G}}^c \cap \mathbf{E}_G^c|/\{p(p-4)\}$, where $p(p-4) = p^2 - p - 3p$ is the number of zero non-diagonal entries in $\mathbf{B}$;
- the rate of correct identification of existence (with the right direction) and non-existence of edges defined as $(|\mathbf{E}_{\hat{G}} \cap \mathbf{E}_G| + |\mathbf{E}_{\hat{G}}^c \cap \mathbf{E}_G^c|)/\{p(p-1)/2\}$; and lastly,
- the Frobenius norm of $\mathbf{B} - \hat{\mathbf{B}}$ divided by the number of off-diagonal entries of $\mathbf{B}$, that is, $\mathrm{trace}\{(\mathbf{B} - \hat{\mathbf{B}})(\mathbf{B} - \hat{\mathbf{B}})^{\mathsf{T}}\}/\{p(p-1)\}$.

The first four metrics are of interest when one is concerned about inference on the graph structure, and the last metric is of interest when one wishes to understand the strength of associations between nodes, and to use the estimated Bayesian network for prediction.

5.3 | Simulation Results

Figure 3 depicts the Monte Carlo (MC) averages across ten graphs of TPR, FDR, specificities, and rates of correct identification of existence/non-existence of directed edges associated with three considered methods when $p = 10$ under two specifications of $\mathbf{\Sigma}_u$. Figure 4 shows the same collection of results when $p = 20$. Across these four metrics, the advantages of the score-based methods pairing with the SIC tuning parameter selector are evident over a wide range of reliability ratio τ, whether the PCD algorithm is used for implementing the corrected score method, or the NPS algorithm is used. The naive likelihood-based method suffers from low TPR, although it is comparable with the score-based methods in terms of specificity. This phenomenon can be explained by the attenuation effect of measurement error on slope parameters estimates in a linear regression model [51, Section 1.1]. More specifically, in

the context of linear regression with additive covariates measurement error, naive estimators of covariate effects tend to attenuate towards zero, which explains the low TPR. Such attenuation effect does not compromise naive estimation of a null covariate effect, which explains the robustness of specificity to measurement error. As a combination of TPR and specificity, the correction rate observed for the naive method is also less affected by measurement error than TPR alone. This robustness is more evident in a sparser graph, such as a graph consisting of $p = 20$ nodes with $3p$ edges when comparing with a graph consisting of $p = 10$ nodes with $3p$ edges. Here, a measure of sparsity of a graph G can be defined as $|\mathbf{E}_G|/\{p(p-1)/2\}$, where $p(p-1)/2$ is the largest number of edges possible for a DAG with p nodes. Finally, even in the absence of measurement error (i.e., with $\tau = 1$ in Figures 3 and 4), the two score-based methods still outperform the likelihood-based method when TPR and correction rate are considered. This implies that the construction of the (unpenalized) objective function plays an important role in network inference.

Figure 5 shows MC medians of the Frobenius norm of $\mathbf{B} - \hat{\mathbf{B}}$ divided by $p(p-1)$. This figure suggests that the PCD algorithm can lead to some numerical instability for the corrected score method, and the NPS algorithm produces more stable regression coefficients estimates that are also less biased than the naive estimates. In fact, between the two score-based methods, the one using the NPS algorithm yields better inference outcomes in all aspects depicted in Figures 3–5 than those resulting from the PCD algorithm. This suggests that there may exist some interaction effect of regression coefficients estimation and cycle elimination procedure on the finite sample performance of a method. Algorithmic properties of different optimization methods in conjunction with different cycle detection and elimination procedures deserve a separate investigation that is beyond the scope of the current study.

5.4 | Inference under a misspecified $\mathbf{\Sigma}_u$

To inspect sensitivity of the proposed methodology to misspecification of $\mathbf{\Sigma}_u$, we implement the score-based methods while assuming a variance-covariance matrix for the measurement error to be $\tilde{\mathbf{\Sigma}}_u$, which can differ from the truth, $\mathbf{\Sigma}_u$, in two ways. In the first type of misspecification, $\tilde{\mathbf{\Sigma}}_u$ and $\mathbf{\Sigma}_u$ share the same structure, with $\tilde{\mathbf{\Sigma}}_u = \tilde{\sigma}_u^2 \mathbf{I}_p$ or $\tilde{\mathbf{\Sigma}}_u = \tilde{\sigma}_u^2 \mathbf{V}_p$, where $\tilde{\sigma}_u^2 = \sigma_u^2 + 0.05k$, for $k = 0, \pm 1, \pm 2, \pm 3$, so that one understates or overstates the severity of error contamination unless when $k = 0$. Using error-contaminated data generated with $\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{I}_p$ or $\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{V}_p$, fixing σ_u^2 at 0.25 to yield a reliability ratio of $\tau = 0.8$ for each node, Figure 6 presents MC averages of TPR, FDR, specificities, and rates of correct identification of directed edges when the two score-based methods are used to infer a graph with $p = 10$ nodes. Overall, misspecifying $\mathbf{\Sigma}_u$ by a scale while keeping the right structure only affects TPR and the correction rate slightly, although overstating error contamination tends to inflate FDR and lower the specificity. The latter phenomenon can be due to overcorrecting the strength of some covariate effects when one overstates the measurement error variance. Parallel results when $p = 20$ are shown in Appendix E of the Supplementary Materials, which tell similar stories. Figure 7 depicts Monte Carlo medians of the Frobenius norm of $\mathbf{B} - \hat{\mathbf{B}}$ divided by $p(p-1)$ in the presence of this type of misspecification. When the graph is sparser, i.e., when $p = 20$ in the current setting, the impact of misspecifying $\mathbf{\Sigma}_u$ is milder when the NPS alogorithm is used than when the PCD algorithm is used. The PCD algorithm suffers from a misspecified $\mathbf{\Sigma}_u$ in terms of covariates effects estimation more when measurement errors are correlated.

In the second type of misspecification, the assumed $\tilde{\mathbf{\Sigma}}_u$ is structurally different from the truth. Figure 8 shows counterpart results of Figure 3 when one mistakenly assumes uncorrelated (or correlated) measurement error when $\mathbf{\Sigma}_u$ indicates correlated (or uncorrelated) measurement error. By comparing patterns observed in Figure 8 with those depicted in Figure 3, both figures for cases with $p = 10$, one can see that TPR and the correction rate are more compromised when one mistakenly assumes uncorrelated measurement error in the presence of correlated measurement error; and it is less harmful to assume correlated measurement errors when measurement errors are actually uncorrelated. Other metrics relating to graph structure estimation are fairly robust to this type of misspecification.

Bias in covariates effects estimation are much more substantial than those under the first type of misspecification. Counterpart results when $p = 20$ are summarized in Appendix E of the Supplementary Materials, which indicate that graph structure estimation is more robust to this form of misspecification when the graph is sparser, despite the bias in covariates effects estimation.

Lastly, instead of assuming a variance-covariance matrix for the measurement error, we estimate it based on four replicate measures of $\mathbf{X}$, denoted by $\mathbf{W}_m$, for $m = 1, 2, 3, 4$, in each MC replicate under each simulation setting described in Section 5.2. The estimator follows that in equation (4.3) in Carroll et al. [50], and is given by $\hat{\mathbf{\Sigma}}_u = (3N)^{-1} \sum_{\ell=1}^{N} \sum_{m=1}^{4} (\mathbf{W}_m[\ell,] - \overline{\mathbf{W}}[\ell,])^t (\mathbf{W}_m[\ell,] - \overline{\mathbf{W}}[\ell,])$, where $\overline{\mathbf{W}} = \sum_{m=1}^{4} \mathbf{W}_m/4$. To infer a graph, we use $\overline{\mathbf{W}}$ as a surrogate of $\mathbf{X}$, and set the corresponding measurement error variance-covariance at $\hat{\mathbf{\Sigma}}_u/4$. With the estimated variance-covariance matrix in place of the truth, Figure 9 demonstrates counterpart summary statistics depicted in Figure 3 regarding graph structure estimation when $p = 10$. It turns out that the score-based method using the PCD algorithm is much less sensitive to this replacement than when the NPS algorithm is used. And the latter somewhat degrades in terms of TPR and the correction rate. These patterns are also observed when the graph is sparer with $p = 20$ (see Figure E.3 in Appendix E in the Supplementary Materials). Covariates effects estimation is fairly robust to this additional estimation of $\mathbf{\Sigma}_u$, as one can see in Figure 10.

It is worth noting that, despite the observed degradation in some aspects in the inference results due to a misspecified variance-covariace matrix for the measurement error, the score-based methods mostly still improve over the naive method. This suggests that there is indeed some gain in acknowledging existence of measurement error and making effort to account for it when drawing inference. Under the assumption that $\mathbf{\Sigma}_u$ is known, both PCD and NPS algorithms exploit the nice partition of $\mathbf{B}$, which contains all parameters to be inferred in our study. This partition of $\mathbf{B}$ is parallel to decomposing a Bayesian network into p regression models, each of which is the model that PCD or NPS deals with at each iteration of the algorithm. One loses such clean partition of the set of unknown parameters if one wants to estimate $\mathbf{\Sigma}_u$, based on replicate data for instance, along with $\mathbf{B}$. Hence, these two algorithms cannot be easily revised to incorporate the estimation of $\mathbf{\Sigma}_u$ when validation data or replicate measures are available. Different objective functions in conjunction with new algorithms for optimization are needed for simultaneous estimation of $\mathbf{\Sigma}_u$ and $\mathbf{B}$.

5.5 | Application to Flow Cytometry Data

Now we return to the application of inferring cellular signaling networks using flow cytometry data. In particular, the flow cytometry data entertained in this section consist of $p = 11$ phosphomolecular measurements from each of $N = 7466$ human immune system cells collected in an experiment described in Sachs et al. [9]. In this experiment, a series of stimulatory cues and inhibitory interventions were imposed, producing the observed data matrix as a mixture of observational data and interventional data for the eleven phosphorylated proteins and phospholipids (see Table 1 in [9]). Shojaie and Michailidis [11] applied a penalized likelihood estimation method with LASSO and ALASSO penalty to infer the signaling network while assuming known ordering of the eleven nodes. Without assuming ordering known, Fu and Zhou [13] applied their likelihood-based penalized estimation method on this data set to infer a directed signaling network using the PCD algorithm, also treating the data as measures of the true nodes. Luo and Zhao [12] viewed the observed data as error-contaminated surrogates of the true protein activity levels, and assumed a normal additive measurement error, with an inverse gamma prior distribution for the measurement error variance (common for all nodes). Neither of the aforementioned methods guarantees that the inferred graph is acyclic.

As in Luo and Zhao [12], we treat the observed phosphomolecular measurements as error-contaminated measures of the true nodes, whose ordering is unknown. Because this data do not contain replicate measures of the same

underlying protein activity level, error variance is not identifiable, even with the normality assumption imposed. We thus follow a widely adopted practice in the measurement error literature in this case, and carry out sensitivity analysis by inferring the underlying network under different assumed variance-covariance matrices for the measurement error. In particular, we first assume correlated measurement error with $\tilde{\mathbf{\Sigma}}_u = (1-\tau)\hat{\mathbf{\Sigma}}_{\mathbf{W}}$, where $\hat{\mathbf{\Sigma}}_{\mathbf{W}}$ is the sample variance-covariance of $\mathbf{W}$ computed using interventional data; we then assume uncorrelated measurement error with $\tilde{\mathbf{\Sigma}}_u = (1-\tau)\mathrm{diag}(\hat{\sigma}^2_{w_1}, \ldots, \hat{\sigma}^2_{w_p})$, where $\hat{\sigma}^2_{w_j}$ is the j^{th} diagonal entry of $\hat{\mathbf{\Sigma}}_{\mathbf{W}}$, for $j = 1, \ldots, p$.

We apply our score-based methods with $\tau = 0.8, 0.9$ in the above assumed $\tilde{\mathbf{\Sigma}}_u$. The computer code for this data analysis along with the data are available in the supplementary materials. Panel (a) in Figure 11 shows a network with directed edges reflecting causal relationships between these nodes that are currently well accepted in the literature. Networks shown in panels (b) and (c) in Figure 11 are provided by two existing works, including the one in Shojaie and Michailidis [11] where the authors used the ALASSO penalty under the assumption that data are free of measurement error with ordering known, and the network from Fu and Zhou [13] assuming error-free data, where the authors used the PCD algorithm to minimize the penalized negative log-likelihood function as in the naive method considered in our empirical study but without the last step of cycle elimination, and with the ALASSO penalty instead of SCAD. Panels (d)–(f) in Figure 11 are estimated networks from our analysis of the data, carried out in the same way as in the simulation experiment, including the naively inferred network and the networks resulting from the score-based methods while assuming correlated measurement error with $\tau = 0.9$. When comparing each of the latter five networks with the consensus graph, the network from Shojaie and Michailidis [11] includes 14 edges in the consensus network among a total of 27 edges in their inferred graph; and there are 8 edges in the consensus network included in the network from Fu and Zhou [13], which also has a total of 27 edges. Among the remaining three estimated graphs (in panels (d)–(f)), the naive method produces a very sparse graph, with merely 8 edges, among which 5 are in the consensus graph; the corrected score method implemented via the PCD algorithm leads to a graph with 27 edges, 9 of which are in the consensus graph; the corrected score method using the NPS algorithm results in a graph with 26 edges, 8 of which are in the consensus graph.

Figure 12 reproduces the two inferred graphs from the score-based methods in panels (e) and (f) of Figure 11, in comparison with counterpart graphs when we assume a nondiagonal $\tilde{\mathbf{\Sigma}}_u$ with $\tau = 0.8$, and those obtained when a diagonal $\tilde{\mathbf{\Sigma}}_u$ with $\tau = 0.9$ is assumed in the score-based methods. The estimated graphs are very similar when changing τ from 0.9 to 0.8 in the assumed nondiagonal $\tilde{\mathbf{\Sigma}}_u$. For example, when the NPS algorithm is used, the graph under the assumption of $\tau = 0.8$ is identical to that obtained under the assumption of $\tau = 0.9$ except that the latter has two additional edges, and both graphs include the same collection of 8 edges in the consensus graph. When we assumed uncorrelated measurement errors with $\tau = 0.9$, the PCD algorithm results in a much denser graph than that resulting from setting $\tau = 0.9$ in the nondiagonal $\tilde{\mathbf{\Sigma}}_u$, with 33 edges, 8 of which are in the consensus graph that are also in the counterpart graph when assuming uncorrelated measurement error.

Using the consensus graph as a gold standard, the above comparisons between all considered networks suggests that, the naive likelihood-based method with cycle elimination can lead to low discovery rate, and the corrected score methods can identify more truly existing causal relationships between nodes. Between the two methods based on the corrected score, the PCD algorithm can result in a higher false discovery rate than the NPS algorithm. Inferences from the score-based methods are more sensitive to the correlation structure of the measurement error than to the magnitude of the error variance.

6 | DISCUSSION

We proposed score-based methods to infer a Bayesian network using error-prone data from interventional experiments. When only observational data are available, the proposed method can be used to infer graphs within a Markov equivalence class [36], since a graph is not identifiable using observational data only but a Markov equivalence class is. A consistent model criterion is also constructed based on the same score function for tuning parameter selection. Besides establishing the consistency in the resulting regression coefficients estimator, we also provide convincing empirical evidence to show that the proposed score-based methods can substantially outperform a naive likelihood-based method that ignores measurement error. And, even in the absence of measurement error in nodes, using a quadratic inference function constructed based on an unbiased score is more preferable than using a likelihood function to formulate a penalized objective function for network estimation.

We exploit Kahn's topological sorting algorithm along with the PCD algorithm or the NPS algorithm to estimate the regression coefficients matrix, which are computationally less burdensome than many search-and-score methods that aim to select a graph from a DAG family of size that grows super-exponentially fast as p grows [65]. Between the PCD algorithm and the NPS algorithm, the latter is computationally much more efficient. For example, in the simulation experiments presented in Section 5 carried out on a Dell Precision M4800 workstation with 512GB serial ATA solid state drive, it typically takes less than half a second for the NPS algorithm to reach to a final estimated graph with $p = 10$ nodes, but it can take around ten seconds for the PCD algorithm. The contrast in implementation time is even more drastic when $p = 20$, where it can take the PCD more than a minute whereas it takes the NPS algorithm one and a half seconds or so to infer a graph. Instead of applying a topological sorting algorithm to check for acyclicity, incorporating the acyclicity constraint in the score function via a smooth characterization of acyclicity as in Zheng et al. [66] may improve numerical efficiency. One computational hurdle remains for the proposed method when p is large is the inversion of a $(p-1)\times(p-1)$ matrix in (9). A model criterion that does not involve the inversion of a large matrix is more desirable in that case.

We assume identically distributed measurement error when formulating the measurement error models in (2)–(4) and constructing the corrected score in (7). This assumption can be violated if the source of measurement error varies across different experimental conditions in a designed experiment. In this case one may revise the score function in (7) accordingly to reflect non-identically distributed measurement error. As an example, let us consider an experiment involving p conditions, and only node j is intervened under condition j, for $j = 1,\ldots,p$. Denote by $\bar{O}_j$ the complement of the index set associated with observational data for node j, O_j. In other words, $\bar{O}_j$ is the index set associated with interventional data for node j, for $j = 1,\ldots,p$. Suppose that the N rows of $\mathbf{U}$ in (2) are not identically distributed p-dimensional Gaussian measurement errors; instead, $\mathbf{U}[\bar{O}_j,]$ consists of n_j realizations from $N_p(\mathbf{0}, \mathbf{\Sigma}_u^{(j)})$, for $j = 1,\ldots,p$. This is to assume that, within the same experimental condition, measurement errors are i.i.d., but across different conditions, measurement errors may not share the same variance-covariance matrix. To reflect this assumption regarding $\mathbf{U}$, the score function in (7) should be replaced by $\mathbf{\Psi}_j(\mathbf{B}_j) = \sum_{j'=1}^{p}\sum_{\ell\in O_j\cap\bar{O}_{j'}}\{(\mathbf{W}[\ell,j]-\mathbf{W}[\ell,-j]\mathbf{B}_j)\mathbf{W}[\ell,-j]^t + \mathbf{\Sigma}_u^{(j')}[-j,-j]\mathbf{B}_j\}$. If the normality assumption in measurement error is violated, the corrected score in (7) may not be an unbiased score. Constructing score functions that are robust to the normality assumption and also account for measurement error is a follow-up research direction. This is also the direction one can follow to relax the linearity assumption of the regression model in (1).

Supplementary Materials

Supplementary material available online includes Appendix A that provides the proof for Theorem 3.1, Appendices B and C that provide updating formulas for the PCD algorithms in Section 3, Appendix D that provides the proof for the consistency of $\mathrm{SIC}(G)$ stated in Section 4, and Appendix E containing additional simulation results referenced in Section5.4.

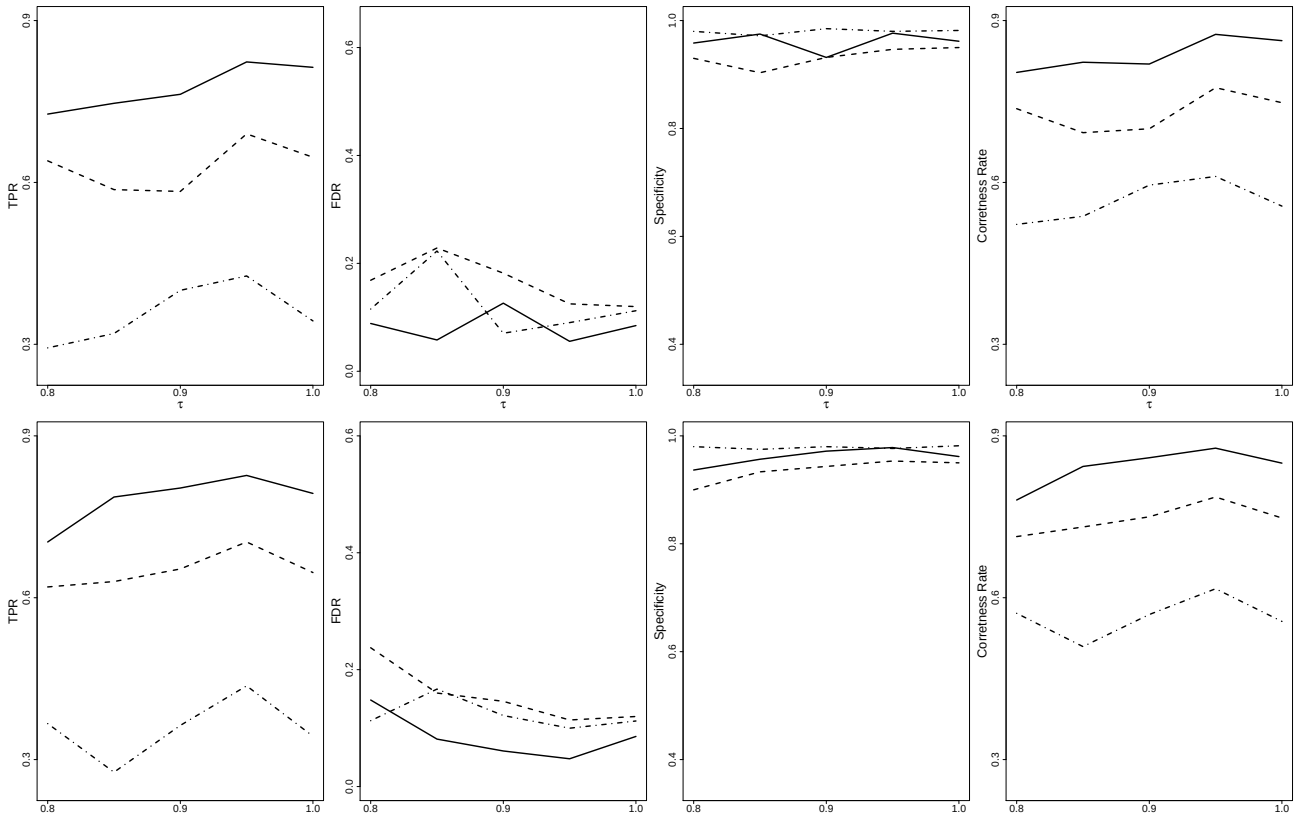

FIGURE 3 Monte Carlo averages of TPR, FDR, specificity, and correctness rate versus the reliability ratio τ across ten graphs with $p = 10$ nodes associated with three methods, the method by Fu and Zhou [13] (dash-dotted lines), corrected score method using PCD algorithm (dashed lines), and corrected score method using NPS algorithm (solid lines), when $\mathbf{\Sigma}_u$ is a diagonal matrix (top panels) and when it is not a diagonal matrix (bottom panels).

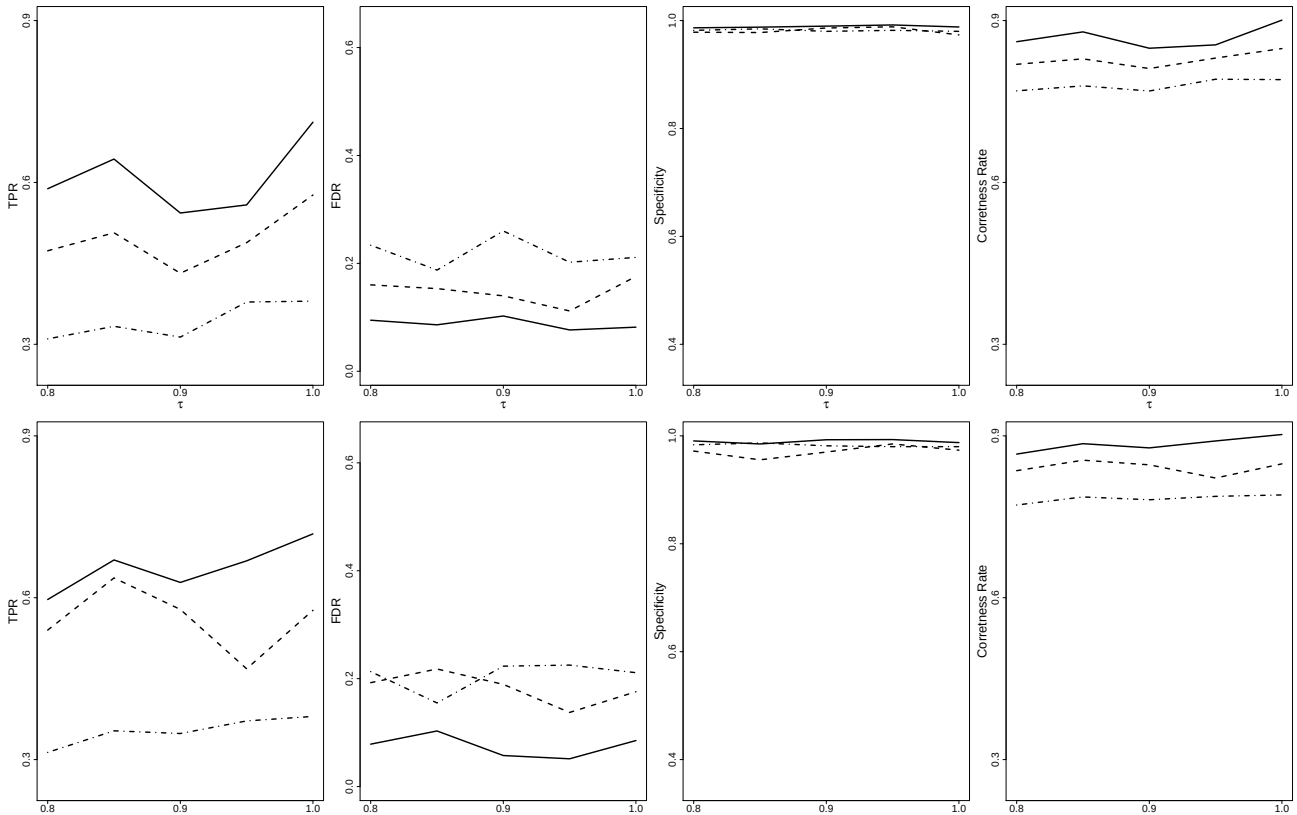

FIGURE 4 Monte Carlo averages of TPR, FDR, specificity, and correctness rate versus the reliability ratio τ across ten graphs with $p = 20$ nodes associated with three methods, the method by Fu and Zhou [13] (dash-dotted lines), corrected score method using PCD algorithm (dashed lines), and corrected score method using NPS algorithm (solid lines), when $\mathbf{\Sigma}_u$ is a diagonal matrix (top panels) and when it is not a diagonal matrix (bottom panels).

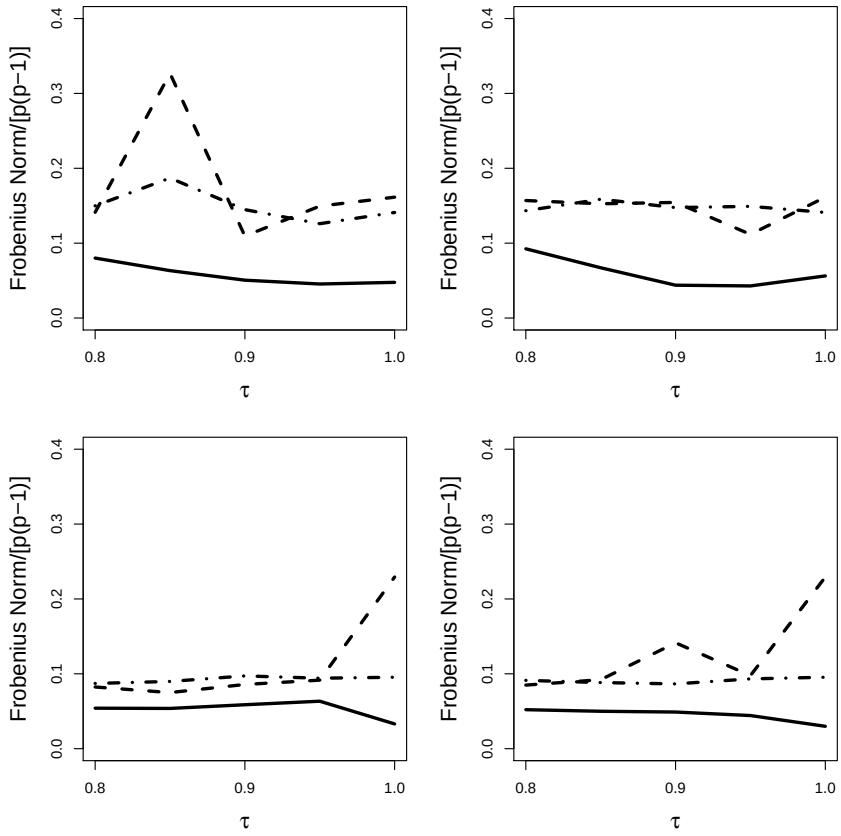

FIGURE 5 Monte Carlo medians of the Frobenius norm of $\mathbf{B} - \hat{\mathbf{B}}$ divided by $p(p-1)$ versus the reliability ratio τ across ten graphs with $p = 10$ nodes (top panels) and $p = 20$ nodes (bottom panels) associated with three methods, the method by Fu and Zhou [13] (dash-dotted lines), corrected score method using PCD algorithm (dashed lines), and corrected score method using NPS algorithm (solid lines), when $\mathbf{\Sigma}_u$ is a diagonal matrix (left panels) and when it is not a diagonal matrix (right panels).

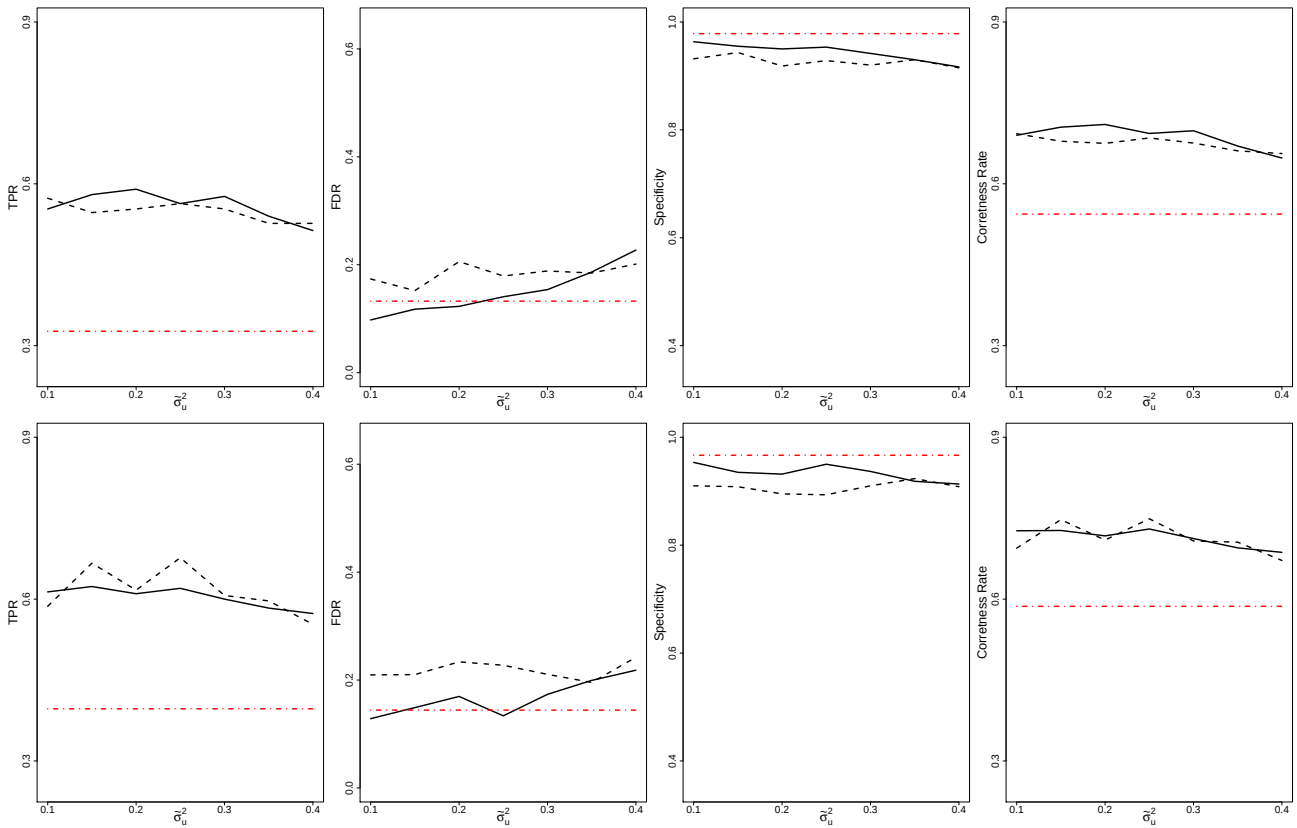

FIGURE 6 Monte Carlo averages of TPR, FDR, specificity, and correctness rate versus $\tilde{\sigma}_u^2$ in the assumed measurement error variance-covariance matrix, $\tilde{\mathbf{\Sigma}}_u = \tilde{\sigma}_u^2 \mathbf{I}_p$ (top panels) and $\tilde{\mathbf{\Sigma}}_u = \tilde{\sigma}_u^2 \mathbf{V}_p$ (bottom panels), across ten graphs with $p = 10$ nodes associated with the corrected score method using PCD algorithm (dashed lines), and the corrected score method using NPS algorithm (solid lines). The red dash-dotted reference line in each panel corresponds to the method by Fu and Zhou [13] that is invariant to the assumed $\tilde{\sigma}_u^2$. The true measurement error variance-covariace metrix is $\mathbf{\Sigma}_u = 0.25\mathbf{I}_p$ (top panels) and $\mathbf{\Sigma}_u = 0.25\mathbf{V}_p$ (lower panels).

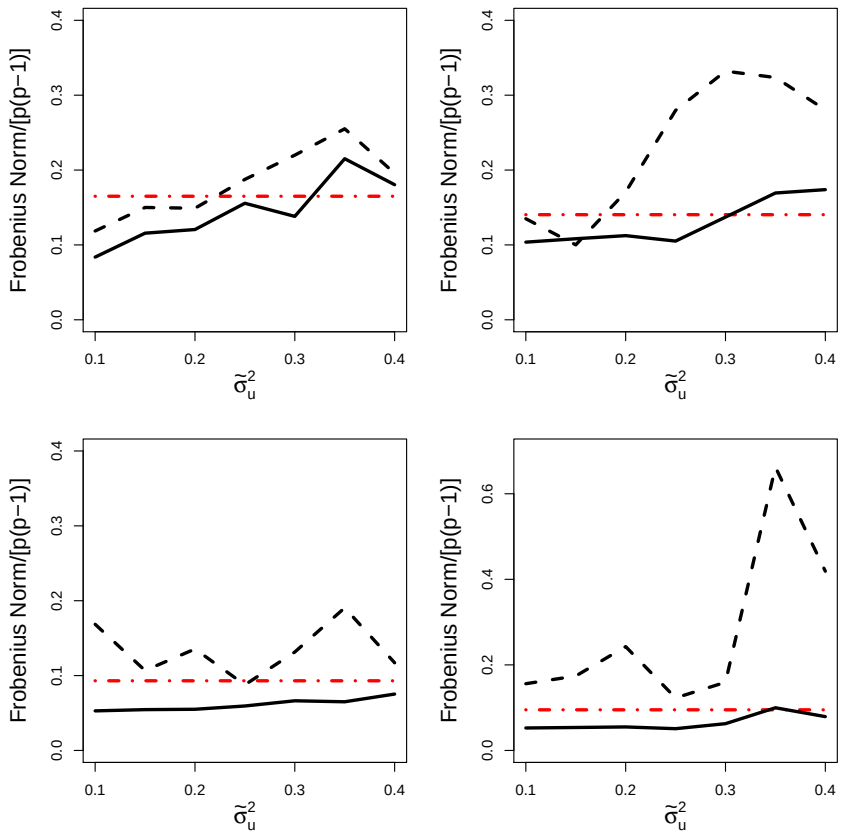

FIGURE 7 Monte Carlo medians of the Frobenius norm of $\mathbf{B} - \hat{\mathbf{B}}$ divided by $p(p-1)$ versus $\tilde{\sigma}^2$ in the assumed measurement error variance-covariance matrix, $\tilde{\mathbf{\Sigma}}_u = \tilde{\sigma}^2\mathbf{I}_p$ (left panels) and $\tilde{\mathbf{\Sigma}}_u = \tilde{\sigma}^2\mathbf{V}_p$ (right panels) across ten graphs with $p = 10$ nodes (top panels) and $p = 20$ nodes (bottom panels) associated with the corrected score method using PCD algorithm (dashed lines) and corrected score method using NPS algorithm (solid lines). The red dash-dotted reference line in each panel corresponds to the method by Fu and Zhou [13] that is invariant to the assumed $\tilde{\sigma}_u^2$. The true measurement error variance-covariace matrix is $\mathbf{\Sigma}_u = 0.25\mathbf{I}_p$ (left panels) and $\mathbf{\Sigma}_u = 0.25\mathbf{V}_p$ (right panels).

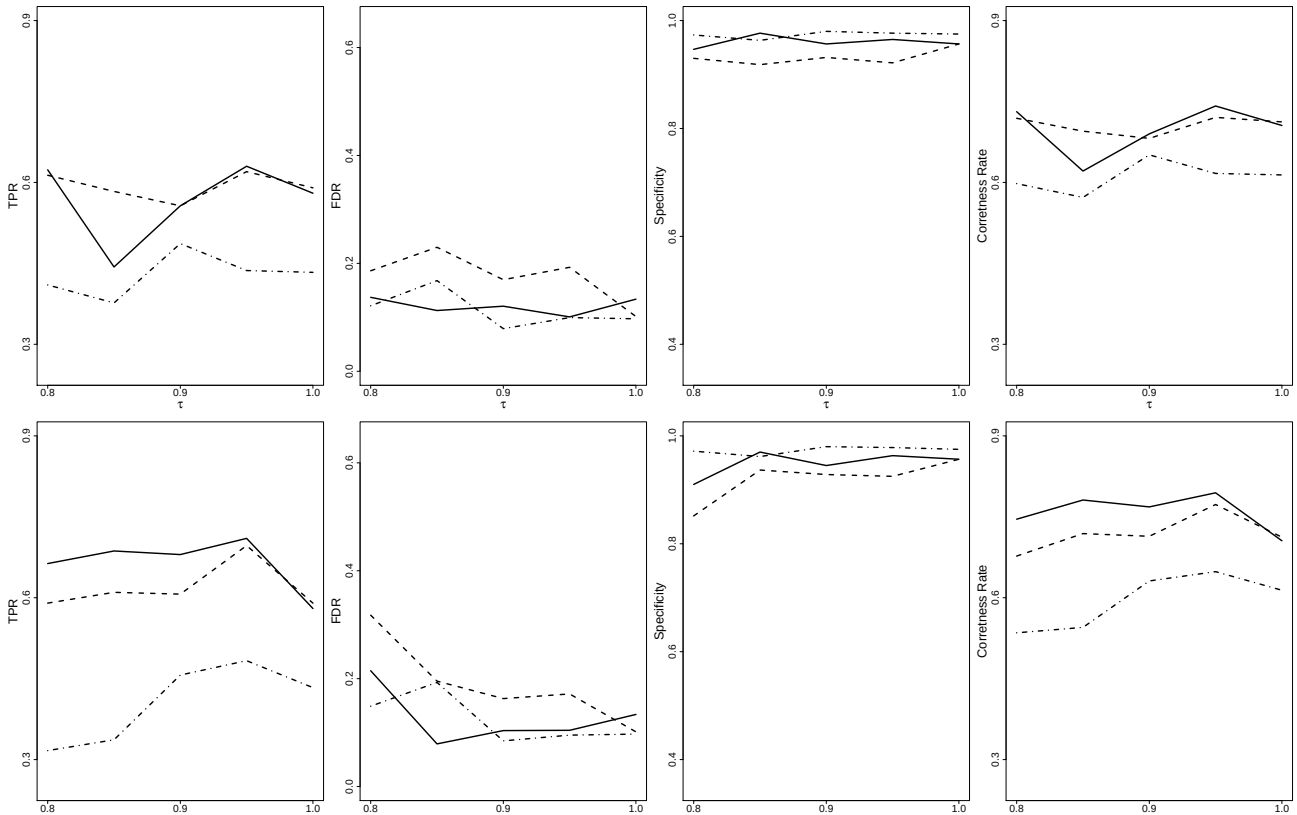

FIGURE 8 Monte Carlo averages of TPR, FDR, specificity, and correctness rate versus the reliability ratio τ across ten graphs with $p = 10$ nodes associated with three methods, the method by Fu and Zhou [13] (dash-dotted lines), corrected score method using PCD algorithm (dashed lines), and corrected score method using NPS algorithm (solid lines), when one assumes $\tilde{\boldsymbol{\Sigma}}_u = \sigma_u^2 \mathbf{I}_p$ while the truth is $\boldsymbol{\Sigma}_u = \sigma_u^2 \mathbf{V}_p$ (top panels) and when one assumes $\tilde{\boldsymbol{\Sigma}}_u = \sigma_u^2 \mathbf{V}_p$ while the truth is $\boldsymbol{\Sigma}_u = \sigma_u^2 \mathbf{I}_p$ (bottom panels).

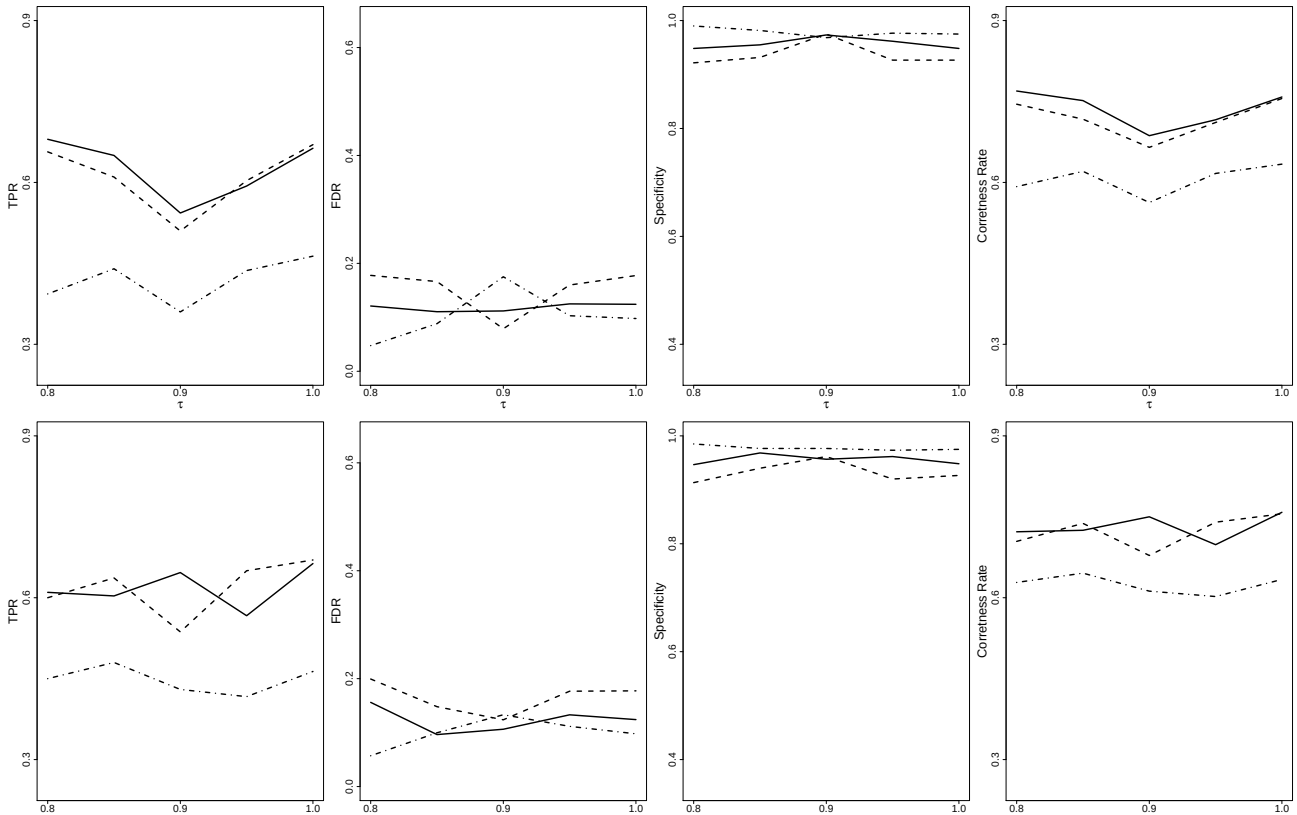

FIGURE 9 Monte Carlo averages of TPR, FDR, specificity, and correctness rate versus the reliability ratio τ across ten graphs with $p = 10$ nodes associated with three methods, the method by Fu and Zhou [13] (dash-dotted lines), corrected score method using PCD algorithm (dashed lines), and corrected score method using NPS algorithm (solid lines), when an estimated variance-covariance matrix for measurement error is used in place of the true $\mathbf{\Sigma}_u$, which is a diagonal matrix (top panels) or when it is not a diagonal matrix (bottom panels).

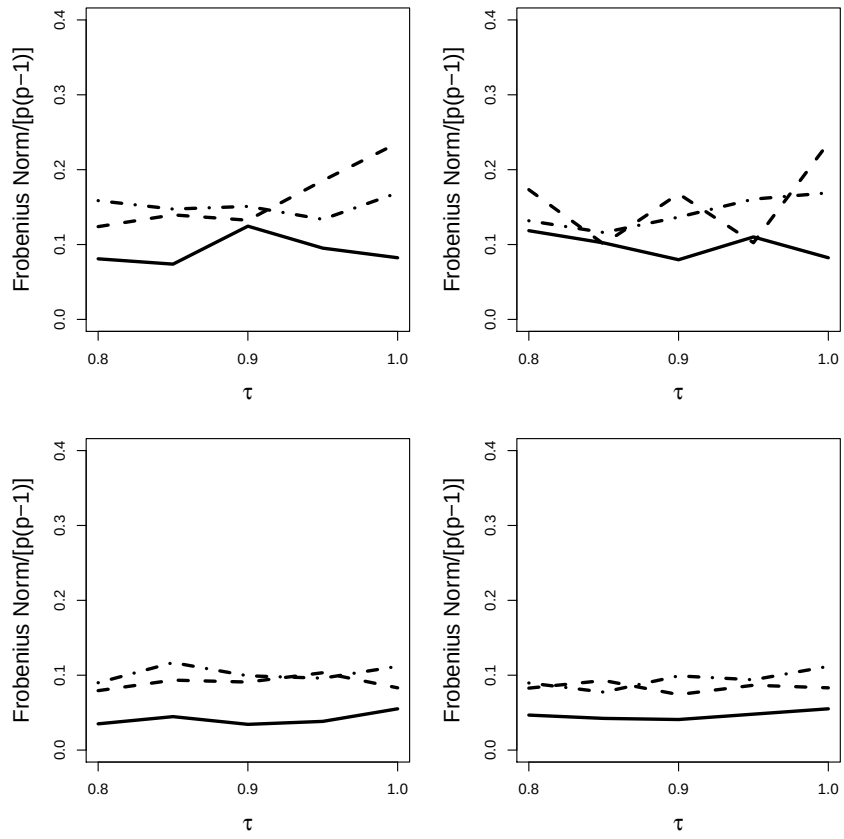

FIGURE 10 Monte Carlo medians of the Frobenius norm of $\mathbf{B} - \hat{\mathbf{B}}$ divided by $p(p-1)$ versus the reliability ratio τ across ten graphs with $p = 10$ nodes (top panels) and $p = 20$ nodes (bottom panels) associated with three methods, the method by Fu and Zhou [13] (dash-dotted lines), corrected score method using PCD algorithm (dashed lines), and corrected score method using NPS algorithm (solid lines), when an estimated variance-covariance matrix for measurement error is used in place of the true $\mathbf{\Sigma}_u$, which is a diagonal matrix (left panels) or when it is not a diagonal matrix (right panels).

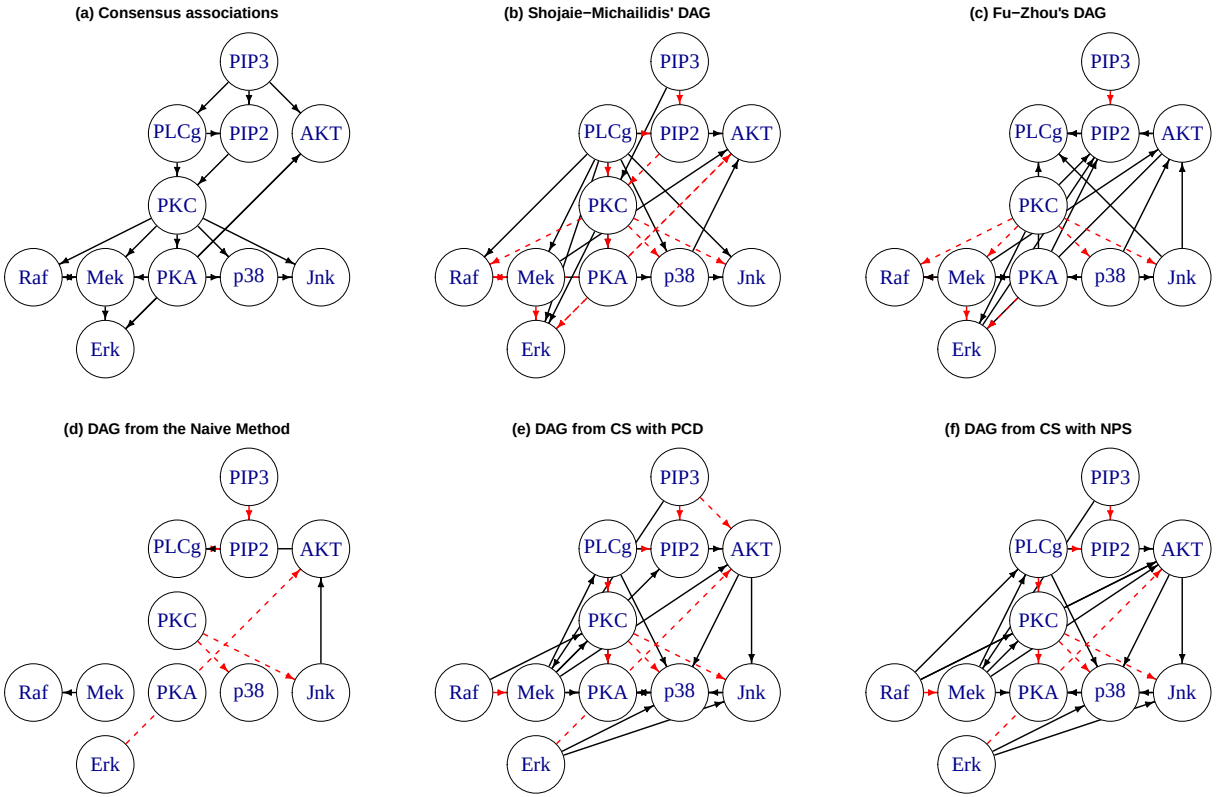

FIGURE 11 Six signaling networks associated with the flow cytometry data set: (a) the consensus graph, (b) the estimated graph from Shojaie and Michailidis [11] assuming ordering known, (c) the estimated graph from Fu and Zhou [13], (d) the estimated graph from the naive method with cycle elimination, (e) the estimated graph from the corrected score method implemented via the PCD algorithm, (f) the estimated graph from the corrected score method implemented via the NPS algorithm. In graphs (b)–(f), the inferred edges in agreement with (a) are highlighted as red dashed edges. In (e) and (f), we assume correlated measurement errors with $\tau = 0.9$ in $\hat{\boldsymbol{\Sigma}}_u = (1-\tau)\hat{\boldsymbol{\Sigma}}_{\mathbf{w}}$.

ACKNOWLEDGEMENTS

We are thankful to the Associate Editor and two anonymous reviewers for their insightful comments and suggestions, which led to a much improved manuscript. Zhang's work is partially supported by National Institutes of Health, NIAID grant R01AI121226.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

references

[1] Jordan JD, Landau EM, Iyengar R. Signaling networks: the origins of cellular multitasking. Cell 2000;103(2):193–200.

[2] Quaranta V, Tyson DR. What lies beneath: looking beyond tumor genetics shows the complexity of signaling networks underlying drug sensitivity. Sci Signal 2013;6(294):pe32–pe32.

[3] Madireddy L, Patsopoulos NA, Cotsapas C, Bos ea Steffan D. A systems biology approach uncovers cell-specific gene regulatory effects of genetic associations in multiple sclerosis. Nature Communications 2019;10:2236.

[4] Eungdamrong NJ, Iyengar R. Modeling cell signaling networks. Biology of the Cell 2004;96(5):355–362.

[5] Kolitz SE, Lauffenburger DA. Measurement and modeling of signaling at the single-cell level. Biochemistry 2012;51(38):7433–7443.

[6] Janes KA, Lauffenburger DA. Models of signalling networks–what cell biologists can gain from them and give to them. J Cell Sci 2013;126(9):1913–1921.

[7] Karamouzis MV, Papavassiliou AG. Tackling the cancer signal transduction "Labyrinth": a combinatorial use of biochemical tools with mathematical models will enhance the identification of optimal targets for each molecular defect. Cancer 2014;120(3):316–322.

[8] Jensen FV. An introduction to Bayesian networks. UCL press London; 1996.

[9] Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science 2005;308(5721):523–529.

[10] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 2008;9(3):432–441.

[11] Shojaie A, Michailidis G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika 2010;97(3):519–538.

[12] Luo R, Zhao H. Bayesian hierarchical modeling for signaling pathway inference from single cell interventional data. The Annals of Applied Statistics 2011;5(2A):725–745.

[13] Fu F, Zhou Q. Learning sparse causal Gaussian networks with experimental intervention: regularization and coordinate descent. Journal of the American Statistical Association 2013;108(501):288–300.

[14] Roederer M. Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats. Cytometry: The Journal of the International Society for Analytical Cytology 2001;45(3):194–205.

[15] Petrunkina A, Harrison R. Systematic misestimation of cell subpopulations by flow cytometry: a mathematical analysis. Theriogenology 2010;73(7):839–847.

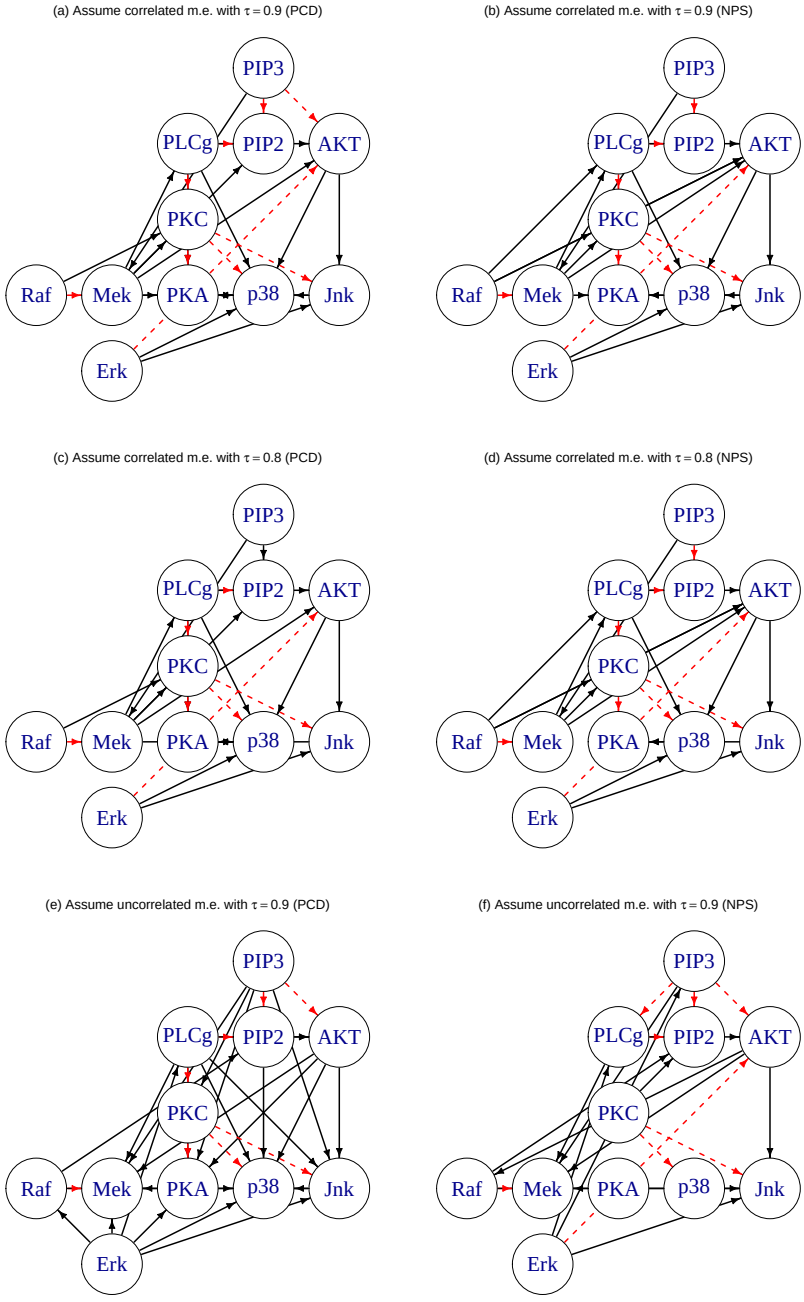

FIGURE 12 Estimated signaling networks from the corrected score method, implemented using the PCD algorithm (left panels) and the NPS algorithm (right panels), respectively, under the assumed variance-covariance matrix for the measurement error given by $\tilde{\boldsymbol{\Sigma}}_u = (1-\tau)\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}$ with $\tau = 0.9$ in (a) and (b), with $\tau = 0.8$ in (c) and (d), and $\tilde{\boldsymbol{\Sigma}}_u = (1-\tau)\mathrm{diag}(\hat{\sigma}^2_{w_1}, \ldots, \hat{\sigma}^2_{w_p})$ with $\tau = 0.9$ in (e) and (f). The inferred edges in agreement with the consensus graph are highlighted as red dashed edges.

[16] Tiberi S, Walsh M, Cavallaro M, Hebenstreit D, Finkenstädt B. Bayesian inference on stochastic gene transcription from flow cytometry data. Bioinformatics 2018;34(17):i647–i655.

[17] Galbusera L, Bellement-Theroue G, Urchueguia A, Julou T, van Nimwegen E. Using fluorescence flow cytometry data for single-cell gene expression analysis in bacteria. PloS one 2020;15(10):e0240233.

[18] Wang DJ, Shi X, McFarland DA, Leskovec J. Measurement error in network data: A re-classification. Social Networks 2012;34(4):396–409.

[19] Li R, Yu J, Zhang S, Bao F, Wang P, Huang X, et al. Bayesian network analysis reveals alterations to default mode network connectivity in individuals at risk for Alzheimer's disease. PLoS One 2013;8(12):e82104.

[20] Ma P, Castillo-Davis CI, Zhong W, Liu JS. A data-driven clustering method for time course gene expression data. Nucleic Acids Research 2006;34(4):1261–1269.

[21] Strimmer K. Modeling gene expression measurement error: a quasi-likelihood approach. BMC bioinformatics 2003;4(1):1–10.

[22] Evans C, Hardin J, Stoebel D. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. arXiv:160900959 2016;.

[23] Neapolitan RE. Probabilistic reasoning in expert systems: theory and algorithms. CreateSpace Independent Publishing Platform; 2012.

[24] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann; 2014.

[25] Lauritzen SL. Graphical models. Clarendon Press; 1996.

[26] Edwards D. Introduction to graphical modelling. Springer Science & Business Media; 2012.

[27] Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. Journal of Computational Biology 2000;7(3-4):601–620.

[28] Suzuki J. A construction of Bayesian networks from databases based on an MDL principle. In: Uncertainty in Artificial Intelligence Elsevier; 1993. p. 266–273.

[29] Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning 1995;20(3):197–243.

[30] Xiang Y, Wong SKM, Cercone N. A microscopic study of minimum entropy search in learning decomposable Markov networks. Machine Learning 1997;26(1):65–92.

[31] Friedman N, Nachman I, Peér D. Learning Bayesian network structure from massive datasets: the sparse candidate algorithm. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence Morgan Kaufmann Publishers Inc.; 1999. p. 206–215.

[32] Chickering DM. Learning equivalence classes of Bayesian-network structures. Journal of Machine Learning Research 2002;2(Feb):445–498.

[33] Moore A, Wong WK. Optimal reinsertion: A new search operator for accelerated and more accurate Bayesian network structure learning. In: Proceedings of the Twentieth International Conference on Machine Learning, vol. 3; 2003. p. 552–559.

[34] Bartlett M, Cussens J. Integer linear programming for the Bayesian network structure learning problem. Artificial Intelligence 2017;244:258–271.

[35] Correia AH, Cussens J, de Campos CP. On Pruning for Score-Based Bayesian Network Structure Learning. arXiv preprint arXiv:190509943 2019;.

[36] Andersson SA, Madigan D, Perlman MD. A characterization of Markov equivalence classes for acyclic digraphs. The Annals of Statistics 1997;25(2):505–541.

[37] Li C, Shen X, Pan W. Likelihood ratio tests for a large directed acyclic graph. Journal of the American Statistical Association 2019;p. 1–16.

[38] Alon N, Yuster R, Zwick U. Color-coding. Journal of the ACM (JACM) 1995;42(4):844–856.

[39] Van de Geer S, Bühlmann P, et al. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. The Annals of Statistics 2013;41(2):536–567.

[40] Gu J, Fu F, Zhou Q. Penalized estimation of directed acyclic graphs from discrete data. Statistics and Computing 2019;29(1):161–176.

[41] Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. Social science computer review 1991;9(1):62–72.

[42] Spirtes P, Glymour CN, Scheines R. Causation, prediction, and search. MIT press; 2000.

[43] Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning 2006;65(1):31–78.

[44] Friedman N, Koller D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. Machine Learning 2003;50(1-2):95–125.

[45] Eaton D, Murphy K. Bayesian structure learning using dynamic programming and MCMC. arXiv:12065247 2012;.

[46] Ellis B, Wong WH. Learning causal Bayesian network structures from experimental data. Journal of the American Statistical Association 2008;103(482):778–789.

[47] Ye Q, Amini AA, Zhou Q. Optimizing regularized Cholesky score for order-based learning of Bayesian networks. arXiv preprint arXiv:190412360 2019;.

[48] Hauser A, Bühlmann P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. Journal of Machine Learning Research 2012;13(Aug):2409–2464.

[49] Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2016;78(5):947–1012.

[50] Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. Chapman & Hall/CRC; 2006.

[51] Fuller WA. Measurement error models, vol. 305. John Wiley & Sons; 2009.

[52] Grace YY. Statistical analysis with measurement error or misclassification. Springer; 2016.

[53] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological) 1996;58(1):267–288.

[54] Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 2006;101(476):1418–1429.

[55] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 2001;96(456):1348–1360.

[56] Aragam B, Zhou Q. Concave penalized estimation of sparse Gaussian Bayesian networks. Journal of Machine Learning Research 2015;16(1):2273–2328.

[57] Nakamura T. Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. Biometrika 1990;77(1):127–137.

[58] Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. Biometrika 2000;87(4):823–836.

[59] Hansen LP. Large sample properties of generalized method of moments estimators. Econometrica 1982;50:1029–1054.

[60] Kahn AB. Topological sorting of large networks. Communications of the ACM 1962;5(11):558–562.

[61] Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to algorithms second edition. The Knuth-Morris-Pratt Algorithm, year 2001;.

[62] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning 2011;3(1):1–122.

[63] Huang X, Zhang H. Variable selection in linear measurement error models via penalized score functions. Journal of Statistical Planning and Inference 2013;143(12):2101–2111.

[64] Ma Y, Li R. Variable selection in measurement error models. Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability 2010;16(1):274.

[65] Robinson. Counting labeled acyclic digraphs. In: New Directions in the Theory of Graphs Academic Press; 1973. p. 239–273.

[66] Zheng X, Aragam B, Ravikumar PK, Xing EP. DAGs with NO TEARS: Continuous optimization for structure learning. In: Advances in Neural Information Processing Systems; 2018. p. 9472–9483.