Received: 15 October 2019 | Revised: 20 April 2021 | Accepted: 25 April 2021

DOI: 10.1002/sim.9028

RESEARCH ARTICLE

Statistics in Medicine WILEY

Penalized estimation of the Gaussian graphical model from data with replicates

Wessel N. van Wieringen[1,2] | **Yao Chen**[3]

[1]Department of Epidemiology and Data Science, Amsterdam Public Health Research Institute, Amsterdam UMC, location VUmc, Amsterdam, The Netherlands

[2]Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[3]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

Correspondence
Wessel N. van Wieringen, Department of Epidemiology and Data Science, Amsterdam Public Health research institute, Amsterdam UMC, location VUmc, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands.
Email: w.vanwieringen@amsterdamumc.nl

Funding information
Euratom research and training programme 2014–2018, Grant/Award Number: 755523

Gaussian graphical models are usually estimated from unreplicated data. The data are, however, likely to comprise signal and noise. These two cannot be deconvoluted from unreplicated data. Pragmatically, the noise is then ignored in practice. We point out the consequences of this practice for the reconstruction of the conditional independence graph of the signal. Replicated data allow for the deconvolution of signal and noise and the reconstruction of former's conditional independence graph. Hereto we present a penalized Expectation-Maximization algorithm. The penalty parameter is chosen to maximize the F-fold cross-validated log-likelihood. Sampling schemes of the folds from replicated data are discussed. By simulation we investigate the effect of replicates on the reconstruction of the signal's conditional independence graph. Moreover, we compare the proposed method to several obvious competitors. In an application we use data from oncogenomic studies with replicates to reconstruct the gene-gene interaction networks, operationalized as conditional independence graphs. This yields a realistic portrait of the effect of ignoring other sources but sampling variation. In addition, it bears implications on the reproducibility of inferred gene-gene interaction networks reported in literature.

KEYWORDS

conditional independence graph, inverse covariance, network, reproducibility, ridge penalty

1 | INTRODUCTION

Gaussian graphical models are used to model (static) molecular networks.[1] These models, and subsequently the network, are learned from omics data. Such data are typically gene expression data that represent the activity of the entities (ie, genes) that constitute the nodes of the network. Data used for the aforementioned purpose are usually acquired as a side product of a clinical or an observational study. Within the context of such studies patients are characterized molecularly once, which is mainly due to financial reasons but also a lack of awareness of the importance of replicates. Consequently, for Gaussian graphical modeling one assumes that only sampling variation is present, ignoring other sources of variation. Here we investigate the consequences of this assumption for the reconstruction of the molecular network.

This author contributed to the work while at the Department of Mathematics, Leiden University, Leiden, The Netherlands.

A Gaussian graphical model is a multivariate normal distribution $\mathcal{N}(\mathbf{0}_p, \mathbf{\Omega}^{-1})$ where $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ is the inverse covariance matrix, henceforth called precision matrix. The specification of the multivariate normal in terms of the precision is due to the fact that the off-diagonal elements of $\mathbf{\Omega}$ contain the information on the conditional (in)dependencies among the variates. A zero off-diagonal element implies that the corresponding variates are conditionally independent, given all other variates, while a nonzero off-diagonal element indicates that there is no such conditional independence. For more on Gaussian graphical models refer to the monographs of Whittaker[2] and Lauritzen.[3]

The parameter of the Gaussian graphical model is usually estimated by means of likelihood maximization. The estimation requires a sample of p-dimensional, independent random vectors $\mathbf{Y}_i$, $i = 1, \dots, n$, from the distribution $\mathcal{N}(\mathbf{0}_p, \mathbf{\Omega}^{-1})$. The maximum likelihood estimator of the precision matrix $\mathbf{\Omega}$ then is the inverse of the sample covariance matrix: $\mathbf{\Omega} = \mathbf{S}^{-1}$, where $\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i \mathbf{Y}_i^{\top}$. When $p > n$, this estimator is not defined as $\mathbf{S}$ is singular. One then resorts to penalized maximum likelihood procedures that define the estimator as the maximizer of the log-likelihood augmented with a penalty term.[4,5]

Both the Gaussian graphical model and the maximum likelihood estimator of its parameter assume that the variation among the random vectors is only due to the sampling from $\mathcal{N}(\mathbf{0}_p, \mathbf{\Omega}^{-1})$ and other sources of variation are absent. In practice, however, such alternative sources of variation, for example, measurement error, are likely to be present in any data as was realized almost a century ago (Shewhart,[6] p. 378): *"An element of chance enters into every measurement; hence every set of measurements is inherently a sample of certain more or less unknown conditions. Even in those few instances where we believe that the objective reality being measured is constant, the measurements of this constant are influenced by chance or unknown causes."*

Translated to the present context, for instance, to acquire an individual's molecular profile the preparation of the sample and the experimentation contribute substantially to the noise in the eventual observation. This dilutes the biological signal present in the data. The presented maximum likelihood estimator of $\mathbf{\Omega}$ then does not estimate the signal precision matrix but a convoluted version of it. The convoluted version may harbor different conditional (in)dependence relations than the original one (examples are given in Section 2).

The direct in-house motivation behind this work stems from an omics study in which only a few samples were replicated. These replications were due to doubts about the quality of some measurements (hybridizations). After closer inspection, it turned out both hybridizations of the few replicated samples were of acceptable quality and were both included in the dataset. Application of standard methods for the estimation of Gaussian graphical models cannot accommodate replicates. To apply such methods would require us to choose one replicate and ignore the other. This felt undesirable and suboptimal. We thus proceeded to modify the aforementioned existing methodology to accommodate replicates. In addition, we seized the opportunity to exploit the inclusion of replicates in the learning of Gaussian graphical models. Here we present the results of this endeavour.

In this work, we investigate how ignoring other sources of variation, such as measurement error, affects the estimated Gaussian graphical model. Hereto we consider studies with a design that is partially replicated. The replicates enable the separation of the sampling variation from that of other causes. Data from such a study are described by a Gaussian graphical model endowed with a "signal+noise" structure. We present its maximum likelihood estimation, in particular high-dimensionally, including the choice of the penalty parameter. In an extensive simulation study, we then investigate the effect of taking into account other sources of variation on the estimation of the signal precision and the related conditional (in)dependence graph. The paper closes with a re-analysis of several Cancer Genome Atlas studies that repeatedly characterized a subset of the included samples transcriptomically by different platforms. The re-analysis demonstrates the effect of ignoring variation due to technical and experimental differences between platforms on the reproducibility of reconstructed molecular networks.

1.1 | Related work

Wainwright[7] considers Gaussian graphical model estimation from corrupted data, which—in light of the Shewhart's quote above—could better be called realistic data. A corrupted observation is formed by the sum of a signal and a noise random variable. Both variables are drawn independently from two different multivariate Gaussian distributions. Wainwright[7] discusses the estimation of the signal's precision matrix, in which knowledge on the noise's precision matrix is assumed to be available from other means than the data at hand (ie, effectively known). This knowledge is then used to correct the sample covariance matrix, from which—through a corrected graphical lasso procedure—the signal's precision matrix is estimated. Hence, replicates are not considered as a means to unravel signal and noise.

The MAQC/SEQC (MicroArray/SEquencing Quality Control) initiatives have used replicates to study the reproducibility of findings reported by studies involving molecular high-throughput techniques.[8-10] Of particular interest here, as that issue is revisited in Section 5, is the study of Zhang et al.[10] In that study, micro-array and RNA-seq platforms are compared with respect to transcriptomic characterization of a certain cancer and clinical endpoint prediction, but not gene-gene interaction network construction.

The reproducibility of reconstructed networks has been studied previously.[11-13] Langfelder et al[11] quantify the evolutionary preservation of networks by comparing networks reconstructed from data of mice and man. Bellot et al[12] carried out a benchmark study of network reconstruction methods comparing their reproducibility between two subsamples of the same experiment. Finally, Vinciotti et al[13] assessed the reproducibility of networks reconstructed from micro-array and RNA-seq platforms, concluding it is poor at the individual edge level but better at an aggregate one. While only the latter study of Vinciotti et al[13] considers a study with replicates, even there the dependency among replicates is not addressed explicitly. Hence, even Vinciotti et al[13] do not separate signal from (technical) noise.

2 | EXPERIMENT, DATA, AND MODEL

Consider an unstructured observational study with certain samples interrogated molecularly multiple times. Let $\mathbf{Y}_{i,k_i}$ be a p-dimensional random variable representing the data resulting from the k_ith replicate, with $k_i = 1, \ldots, K_i$, of this measurement on sample $i = 1, \ldots, n$. We model the data from the described study by an additive model: $\mathbf{Y}_{i,k_i} = \mathbf{Z}_i + \boldsymbol{\varepsilon}_{i,k_i}$. In this model $\mathbf{Z}_i$ can be thought of as the signal present in sample i, while $\boldsymbol{\varepsilon}_{i,k_i}$ represents the noise in the k_ith replicate of sample i. We assume the signal and error both to follow a multivariate normal distribution but with different covariance matrices (as specified by the inverse precision matrices): $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Omega}_z^{-1})$ and $\boldsymbol{\varepsilon}_{ik} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Omega}_\varepsilon^{-1})$, respectively. Additionally, we take the signals and errors to be independent, in the sense that $\mathbf{Z}_i \perp\!\!\!\perp \boldsymbol{\varepsilon}_{ik}$, $\mathbf{Z}_{i_1} \perp\!\!\!\perp \mathbf{Z}_{i_2}$ for $i_1 \neq i_2$, and $\boldsymbol{\varepsilon}_{i,k_1} \perp\!\!\!\perp \boldsymbol{\varepsilon}_{i,k_2}$ for $k_1 \neq k_2$. Thus, $\mathbf{Y}_{ik} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Omega}_z^{-1} + \boldsymbol{\Omega}_\varepsilon^{-1})$ with the following marginal and conditional (in)dependence relations: $\text{Cov}(\mathbf{Y}_{i_1,k_1}, \mathbf{Y}_{i_2,k_2}) = \mathbf{0}_{pp}$ for $i_1 \neq i_2$, $\text{Cov}(\mathbf{Y}_{i,k_1}, \mathbf{Y}_{i,k_2}) = \boldsymbol{\Omega}_z^{-1}$ for $k_1 \neq k_2$, $\text{Cov}(\mathbf{Y}_{i,k_1}, \mathbf{Y}_{i,k_2} \mid \mathbf{Z}_i) = \mathbf{0}_{pp}$ for $k_1 \neq k_2$, and $\text{Cov}(\mathbf{Y}_{i,k_1}, \mathbf{Y}_{i,k_2} \mid \mathbf{Z}_i) = \boldsymbol{\Omega}_\varepsilon^{-1}$ for $k_1 = k_2$.

The above simple "signal+noise" model enables us to illustrate the effect of only taking sampling variation into account when estimating a Gaussian graphical model. In the presence of other sources of variation, as captured by the parameter $\boldsymbol{\Omega}_\varepsilon^{-1}$, one ought to infer the conditional independencies from $\boldsymbol{\Omega}_z$. Common practice, however, bases this inference on the inverse of $\boldsymbol{\Omega}_y^{-1} = \boldsymbol{\Omega}_z^{-1} + \boldsymbol{\Omega}_\varepsilon^{-1}$. This leads to false positively and false negatively inferred edges. To see this consider a numerical example where $\boldsymbol{\Omega}_\varepsilon^{-1} = \mathbf{I}_{33}$ and

$$\boldsymbol{\Omega}_z^{-1} = \begin{pmatrix} 3 & -1 & 2 \\ -1 & 3 & -2 \\ 2 & -2 & 4 \end{pmatrix}.$$

Then, $(\boldsymbol{\Omega}_z)_{1,2} = 0$ but $[(\boldsymbol{\Omega}_z^{-1} + \boldsymbol{\Omega}_\varepsilon^{-1})^{-1}]_{1,2} \neq 0$. Here ignorance of all sources but sampling variation induces a false-positive edge. A numeric example for the opposite, a false-negative edge, is also easily constructed. The difference in the conditional independence graphs inferred from $\boldsymbol{\Omega}_y$ and $\boldsymbol{\Omega}_z$ can be quantified more generally. Hereto use the result of Miller[14] on the inverse of a sum of two matrices to write $\boldsymbol{\Omega}_y$ in terms of $\boldsymbol{\Omega}_z$:

$$\boldsymbol{\Omega}_y = (\boldsymbol{\Omega}_z^{-1} + \boldsymbol{\Omega}_\varepsilon^{-1})^{-1} = \boldsymbol{\Omega}_z - (\mathbf{I}_{pp} + \boldsymbol{\Omega}_z \boldsymbol{\Omega}_\varepsilon^{-1})^{-1} \boldsymbol{\Omega}_z \boldsymbol{\Omega}_\varepsilon^{-1} \boldsymbol{\Omega}_z.$$

Hence, nonzero off-diagonal elements of $(\mathbf{I}_{pp} + \boldsymbol{\Omega}_z \boldsymbol{\Omega}_\varepsilon^{-1})^{-1} \boldsymbol{\Omega}_z \boldsymbol{\Omega}_\varepsilon^{-1} \boldsymbol{\Omega}_z$ reveal differences in the strength of the edges of the conditional independence graphs inferred from observation and signal precision matrices.

To write down the likelihood of the data under the specified model, the following lemma, which is a generalization of the result presented in the appendix A of Riebler et al,[15] is needed. It specifies the elements of the inverse of the joint covariance matrix of the vector of replicates of a sample.

Lemma 1. *Let* $\mathbf{Y}$ *be a multivariate normal random variable partitioned in* K *equally sized blocks as* $\mathbf{Y} = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, \ldots, \mathbf{Y}_K^\top)^\top$ *with covariance matrix* $\boldsymbol{\Omega}_y^{-1} = \mathbf{1}_{KK} \otimes \boldsymbol{\Omega}_z^{-1} + \mathbf{I}_{KK} \otimes \boldsymbol{\Omega}_\varepsilon^{-1}$, *with* $\boldsymbol{\Omega}_z, \boldsymbol{\Omega}_\varepsilon \in \mathcal{S}_{++}^p$, *that is, both* $p \times p$ *dimensional, symmetric and positive definite matrices. Then, its covariance matrix has determinant* $|\boldsymbol{\Omega}_\varepsilon|^{-K+1} |K\boldsymbol{\Omega}_z^{-1} + \boldsymbol{\Omega}_\varepsilon^{-1}|$ *while the blocks of its inverse* *equal:*

$$\begin{aligned}(\mathbf{\Omega}_y)_{k,k'} &= K^{-1}[(K\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_\varepsilon^{-1})^{-1} - \mathbf{\Omega}_\varepsilon],\\ (\mathbf{\Omega}_y)_{k,k} &= K^{-1}[(K\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_\varepsilon^{-1})^{-1} + (K-1)\mathbf{\Omega}_\varepsilon],\end{aligned}$$

for $k, k' = 1, \ldots, K$ *and* $k \neq k'$. *Moreover, the blocks of* $\mathbf{\Omega}_y$ *satisfy:* $(\mathbf{\Omega}_y)_{k,k} - (\mathbf{\Omega}_y)_{k,k'} = \mathbf{\Omega}_\varepsilon$ *and* $(\mathbf{\Omega}_y)_{k,k} + (K-1)(\mathbf{\Omega}_y)_{k,k'} = (K\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_\varepsilon^{-1})^{-1}$.

Proof. The determinant identity follows from the factorization:

$$\mathbf{1}_{KK} \otimes \mathbf{\Omega}_z^{-1} + \mathbf{I}_{KK} \otimes \mathbf{\Omega}_\varepsilon^{-1} = (\mathbf{I}_{KK} \otimes \mathbf{\Omega}_\varepsilon^{-1})(\mathbf{1}_{KK} \otimes \mathbf{\Omega}_\varepsilon \mathbf{\Omega}_z^{-1} + \mathbf{I}_{KK} \otimes \mathbf{I}_{pp}),$$

the determinant of a Kronecker product (cf, section 16.3.e of Harville[16]), the specifics of the eigenvalues of $\mathbf{1}_{KK}$, and the use of well-known results from standard linear algebra on eigen-decompositions and determinants. Furthermore, the inverse is verified by use of straightforward linear algebra. Finally, the relations for the blocks of the inverse are immediate from these analytic expressions. ■

The loglikelihood of the data can now, when invoking Lemma 1 and some algebraic manipulations, be formulated as:

$$\begin{aligned}\sum_{i=1}^{n} \log[P(\mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,K_i})] \propto \sum_{i=1}^{n} \big\{&(K_i - 1)\log(|\mathbf{\Omega}_\varepsilon|) - \log(|K_i\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_\varepsilon^{-1}|)\\ &- (K_i - 1)\mathrm{tr}(\mathbf{\Omega}_\varepsilon \mathbf{S}_{\varepsilon,i}) - \mathrm{tr}[(K_i\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_\varepsilon^{-1})^{-1}\mathbf{S}_{y,i}]\big\},\end{aligned}$$

where $\mathbf{S}_{\varepsilon,i} = (K_i - 1)^{-1}\left(\sum_{k=1}^{K_i} \mathbf{Y}_{i,k}\mathbf{Y}_{i,k}^\top - K_i^{-1}\sum_{k,k'=1}^{K_i} \mathbf{Y}_{i,k}\mathbf{Y}_{i,k'}^\top\right)$ and $\mathbf{S}_{y,i} = K_i^{-1}\sum_{k,k'=1}^{K_i} \mathbf{Y}_{i,k}\mathbf{Y}_{i,k'}^\top$.

3 | ESTIMATION

We estimate the parameters $\mathbf{\Omega}_z$ and $\mathbf{\Omega}_\varepsilon$ by means of likelihood maximization. Its maximizer is found by means of the EM algorithm,[17] an iterative procedure that alternates between the so-called E- and M-steps. The procedure starts from initial parameter estimates. In the E-step, or Expectation step, sufficient statistics for the estimation of the parameters are obtained. In the M-step, or Maximization step, the parameter estimates are updated by means of (complete) likelihood maximization, given the data and the acquired sufficient statistics.

The E-step produces sufficient statistics for the distribution of the unobserved $\mathbf{Z}_i$. As the $\mathbf{Z}_i$ follow a multivariate normal distribution, the sufficient statistics are the sample versions of its first two moments. But as the $\mathbf{Z}_i$ are unobserved, these are replaced by the expectation of these two sample moments conditional on the data using the current parameter estimates. These expectations are found from the joint distribution of $(\mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{K_i,1}, \mathbf{Z}_i)$. This distribution is a zero-centered multivariate normal distribution with covariance matrix:

$$\begin{pmatrix} \mathbf{1}_{K_iK_i} \otimes \mathbf{\Omega}_z^{-1} + \mathbf{I}_{K_iK_i} \otimes \mathbf{\Omega}_\varepsilon^{-1} & \mathbf{1}_{K_i} \otimes \mathbf{\Omega}_z^{-1} \\ \mathbf{1}_{K_iK_i}^\top \otimes \mathbf{\Omega}_z^{-1} & \mathbf{\Omega}_z^{-1} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{K_iK_i} \otimes \mathbf{\Omega}_\varepsilon & -\mathbf{1}_{K_i} \otimes \mathbf{\Omega}_\varepsilon \\ -\mathbf{1}_{K_i}^\top \otimes \mathbf{\Omega}_\varepsilon & \mathbf{\Omega}_z + K_i\mathbf{\Omega}_\varepsilon \end{pmatrix}^{-1}.$$

The inverse on the right-hand side follows from the analytic expression of the inverse of a 2×2 block matrix (theorem 8.5.11 of Harville[16]) in combination with Lemma 1 and the fact that its determinant equals $|\mathbf{\Omega}_\varepsilon|^{-K_i}|\mathbf{\Omega}_z|^{-1}$, which is immediate from theorem 13.3.8 of Harville.[16] Then, using theorem 2.5.1 of Anderson[18] that provides an analytic expression of the conditional distribution of a subset of variates given the others, the aforementioned conditional expectations of the sufficient statistics are:

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i) &= \mathbb{E}(\mathbf{Z}_i \,|\, \mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,K_i}; \mathbf{\Omega}_z, \mathbf{\Omega}_\varepsilon) = \mathbf{\Omega}_z^{-1}(K_i\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_\varepsilon^{-1})^{-1}\sum_{k=1}^{K_i}\mathbf{Y}_{i,k},\\ \mathrm{Var}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i) &= \mathrm{Var}(\mathbf{Z}_i \,|\, \mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,K_i}; \mathbf{\Omega}_z, \mathbf{\Omega}_\varepsilon) = (\mathbf{\Omega}_z + K_i\mathbf{\Omega}_\varepsilon)^{-1}.\end{aligned}$$

In the display above the right-hand side of the first conditional moment is obtained by means of the result presented in Lemma 1 and the second moment follows then from the Inverse Variance lemma (proposition 5.7.3 of Whittaker[2]). These moments need to be evaluated for each sample i, which involves the inverse of $p \times p$-dimensional matrices that depends on K_i. Computationally, it is then most efficient to evaluate these moments of samples with an identical number of replicates as a group such that redundant inversions are avoided. Finally, for use in the M-step these sufficient statistics are evaluated by plugging in the current estimates of the precision matrices.

The M-step finds updates of the parameter estimates, given the estimates of the $\mathbf{Z}_i$ obtained in the E-step, through maximization of the so-called complete likelihood, which is the joint likelihood of $\{(\mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,K_1}, \mathbf{Z}_i)\}_{i=1}^n$. The latter equals:

$$\prod_{i=1}^{n} P(\mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,K_i}, \mathbf{Z}_i) = \prod_{i=1}^{n} P(\mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,K_i} \,|\, \mathbf{Z}_i)\, P(\mathbf{Z}_i) \;=\; \prod_{i=1}^{n} P(\mathbf{Z}_i) \prod_{k=1}^{K_i} P(\mathbf{Y}_{i,k} \,|\, \mathbf{Z}_i).$$

Take the logarithm and obtain the complete log-likelihood:

$$n[\log(|\mathbf{\Omega}_z|) - \mathrm{tr}(\mathbf{\Omega}_z \tilde{\mathbf{S}}_z)] + [\log(|\mathbf{\Omega}_\varepsilon|) - \mathrm{tr}(\mathbf{\Omega}_\varepsilon \tilde{\mathbf{S}}_\varepsilon)] \sum_{i=1}^{n} K_i, \tag{1}$$

where

$$\tilde{\mathbf{S}}_z = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i^\top \quad \text{and} \quad \tilde{\mathbf{S}}_\varepsilon = \frac{1}{\sum_{i=1}^{n} K_i} \sum_{i=1}^{n} \sum_{k=1}^{K_i} (\mathbf{Y}_{i,k} - \mathbf{Z}_i)(\mathbf{Y}_{i,k} - \mathbf{Z}_i)^\top. \tag{2}$$

For the expectation of the complete log-likelihood simply replace $\tilde{\mathbf{S}}_z$ and $\tilde{\mathbf{S}}_\varepsilon$ by their expectations with respect to the $\mathbf{Z}_i \,|\, \mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,K_i}; \mathbf{\Omega}_z, \mathbf{\Omega}_\varepsilon$, which are:

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\tilde{\mathbf{S}}_z) &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i) + \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i)[\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i)]^\top, \\
\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\tilde{\mathbf{S}}_\varepsilon) &= \frac{1}{\sum_{i=1}^{n} K_i} \sum_{i=1}^{n} K_i \mathrm{Var}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i) + \frac{1}{\sum_{i=1}^{n} K_i} \sum_{i=1}^{n} \sum_{k=1}^{K_i} [\mathbf{Y}_{i,k} - \mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i)][\mathbf{Y}_{i,k} - \mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i)]^\top,
\end{aligned}$$

respectively. Maximization of the expected log-likelihood can now be done with respect to the two parameters separately. This yields: $\widehat{\mathbf{\Omega}}_z = [\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\tilde{\mathbf{S}}_z)]^{-1}$ and $\widehat{\mathbf{\Omega}}_\varepsilon = [\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\tilde{\mathbf{S}}_\varepsilon)]^{-1}$. In this, the $\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i)$ and $\mathrm{Var}_{\mathbf{Z}|\mathbf{Y}}(\mathbf{Z}_i)$, as obtained in the E-step, are used for the evaluation of the updated estimates.

The EM algorithm applies the E- and M-step iteratively until convergence. Convergence is reached when the log-likelihood does no longer improve much between subsequent iterations. Following Zhu and Melnykov,[19] we operationalize this as the absolute relative change in the complete log-likelihood. Convergence of the algorithm is then warranted by Jensens' inequality, which implies (after some algebra) that an improvement in the complete log-likelihood implies one in the log-likelihood.[17]

Omics studies, from which the gene-gene interaction network is reconstructed, are often undersampled. The resulting high-dimensional situation requires a modification of the loss criterion. A penalty augments the log-likelihood to ensure the existence of a unique and well-defined estimator. Here the ridge penalty, the sum of the square of the elements of the precision matrices each with their own penalty parameter, that is, $\frac{1}{2}\lambda_z \|\mathbf{\Omega}_z\|_F^2 + \frac{1}{2}\lambda_\varepsilon \|\mathbf{\Omega}_\varepsilon\|_F^2$ with $\|\cdot\|_F$ the Frobenius norm, is used. The maximizer of the thus penalized log-likelihood is found by means of a penalized EM algorithm. This is derived as its unpenalized counterpart. Effectively, the penalization leaves the E-step unaffected and leads to a minor modification of the M-step. In the latter now the expectation of the complete log-likelihood (1) augmented with the ridge penalty is maximized with respect to the parameters. This can done per parameter separately and yields (cf, van Wieringen and Peeters[5]), for example:

$$\widehat{\mathbf{\Omega}}_z(\lambda_z) = \left[\frac{1}{2} \tilde{\mathbf{S}}_z + \left(\tilde{\lambda}_z \mathbf{I}_{pp} + \frac{1}{4} \tilde{\mathbf{S}}_z^2 \right)^{1/2} \right]^{-1},$$

where $\tilde{\lambda}_z = \lambda_z/n$. For that of $\widehat{\mathbf{\Omega}}_\varepsilon$ replace λ_z and $\tilde{\mathbf{S}}_z$ by their contourparts $\tilde{\lambda}_\varepsilon = \lambda_\varepsilon(\sum_{i=1}^{n} K_i)^{-1}$ and $\tilde{\mathbf{S}}_\varepsilon$. In the above the ridge penalty may be replaced by the graphical lasso penalty: $\lambda_z\|\mathbf{\Omega}_z\|_1 + \lambda_\varepsilon\|\mathbf{\Omega}_\varepsilon\|_1$. Estimates of $\mathbf{\Omega}_z$ and $\mathbf{\Omega}_\varepsilon$ are then found by a row/column updating scheme.[4]

3.1 | Diagonal $\mathbf{\Omega}_\varepsilon$

We consider the simplification of the model for estimation of signal and error precision matrices in high-dimensions. While the inclusion of replicates in the design enables the separation of sampling variation from that of other sources, it brings about the estimation of $\frac{1}{2}p(p+1)$ additional parameters (compared to the estimation of the $\mathbf{\Omega}_y$ from an unreplicated design). In addition, an extra penalty parameter needs to be chosen. The recovery of conditional independencies is already a challenging task (especially from high-dimensional studies), but it is further hampered by penalization. Penalization tends to shrink the precision's off-diagonal elements more than its diagonal ones and thereby obstructs the deconvolution of the contributions of signal and error to the conditional (in)dependencies.

The simplification of the model may be achieved by the adoption of assumptions on the structure of precision matrices. This is undesirable for the signal precision matrix as it is the (conditional) relations within the signal that are of primary interest. However, it may be acceptable to make such an assumption for $\mathbf{\Omega}_\varepsilon$ as interest is not in the dependencies among the elements of $\boldsymbol{\varepsilon}_i$. Their independence may therefore be a reasonable simplification (which is investigated in Sections 4 and 5.2). This independence assumption corresponds to a diagonal $\mathbf{\Omega}_\varepsilon$ which involves only p parameters.

Incorporation of the diagonality assumption of $\mathbf{\Omega}_\varepsilon$ into the estimation requires only a minor modification to the penalized EM algorithm. In the M-step the complete likelihood (1) is now maximized with respect to $\mathbf{\Omega}_\varepsilon$ by $[\widehat{\mathbf{\Omega}}_\varepsilon]_{jj} = \{[\mathbf{S}_\varepsilon]_{jj}\}^{-1}$ for $j = 1, \dots, p$, leaving the estimate of $\mathbf{\Omega}_z$ unaffected. In particular, the need for penalization of $\mathbf{\Omega}_\varepsilon$ has vanished as the resulting $\widehat{\mathbf{\Omega}}_\varepsilon$ is well-defined by the independence assumption and the positivity of the estimates of its diagonal elements. The gain in computation time bought by this diagonal $\mathbf{\Omega}_\varepsilon$ assumption is investigated in the SM If of Appendix S1.

We illustrate the effect of the diagonal error assumption on the reconstruction of the conditional independence graph. For simplicity, we assume here $K_i = K$ for all i. We then study the limiting behavior, in either n or K, of the M-step's $\widehat{\mathbf{\Omega}}_z$. Write this inverse of this estimator as:

$$\widehat{\mathbf{\Omega}}_z^{-1} = \mathbf{\Omega}_z^{-1}(K\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_\varepsilon^{-1})^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\mathbf{Y}_{i,k}\right)\left(\sum_{k=1}^{K}\mathbf{Y}_{i,k}\right)^\top\right](K\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_\varepsilon^{-1})^{-1}\mathbf{\Omega}_z^{-1} + (\mathbf{\Omega}_z + K\mathbf{\Omega}_\varepsilon)^{-1},$$

where the analytic expressions for the $\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i)$ and $\mathrm{Var}_{\mathbf{Z}|\mathbf{Y}}(\mathbf{Z}_i)$ have been substituted. For a limiting sample size, note that, by the law of large numbers: $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}(\sum_{k_i=1}^{K}\mathbf{Y}_{i,k_1})(\sum_{k_i=1}^{K}\mathbf{Y}_{i,k_1})^\top = K^2\,\mathbf{\Omega}_z^{-1} + K\mathbf{\Omega}_\varepsilon^{-1}$. Substitution of this limit into the preceding display yields, after some linear algebraic manipulations, $\lim_{n\to\infty}\widehat{\mathbf{\Omega}}_z^{-1} = \lim_{n\to\infty}\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\tilde{\mathbf{S}}_z) = \mathbf{\Omega}_z^{-1}$, in which no assumption on the error precision matrix has been made. Should we erroneously have assumed a diagonal error precision matrix, denoted $\mathbf{\Omega}_{\varepsilon,d}$ and temporarily known, the limit becomes:

$$\lim_{n\to\infty}\widehat{\mathbf{\Omega}}_z^{-1} = \mathbf{\Omega}_z^{-1} + K\mathbf{\Omega}_z^{-1}(K\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_{\varepsilon,d}^{-1})^{-1}(\mathbf{\Omega}_\varepsilon^{-1} - \mathbf{\Omega}_{\varepsilon,d}^{-1})(K\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_{\varepsilon,d}^{-1})^{-1}\mathbf{\Omega}_z^{-1}.$$

The second summand characterizes the effect of the diagonal error precision matrix assumption, which vanishes when diagonality is justified. On the other hand, with a fixed sample size n but a large number of replicates K the assumption becomes irrelevant. Put differently, in the M-step of the algorithm $\lim_{K\to\infty}\widehat{\mathbf{\Omega}}_z = [\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\tilde{\mathbf{S}}_z)]^{-1} = (n^{-1}\sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i^\top)^{-1}$ as $\lim_{K\to\infty}\mathbb{E}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i) = \mathbf{Z}_i$ and $\lim_{K\to\infty}\mathrm{Var}_{\mathbf{Z}\,|\,\mathbf{Y}}(\mathbf{Z}_i) = \mathbf{0}_{pp}$. Intuitively, this is evident when $\mathbf{Z}_i$ are estimated by the average of the $\mathbf{Y}_{i,1}, \dots, \mathbf{Y}_{i,K}$. For large K the error thus averages out. Consequently, the diagonal error assumption does not affect the estimate of $\mathbf{Z}_i$, or the associated $\widehat{\mathbf{\Omega}}_z$. For small n and K, simulations revealed (not shown) that the model with a full error precision matrix performs (slightly) better in edge recovery than that with a diagonal one. The fit of the former is generally better than the latter, unless—of course—the error precision matrix is indeed diagonal.

3.2 | Penalty selection

We choose the penalty parameters λ_z and λ_ε for the signal and error precision matrices by means of F-fold cross-validation (with $F \in \mathbb{N}$ such that $2 \leq F \leq n$). This procedure evaluates—for given λ_z and λ_ε—the performance (in some sense) of estimated precision matrices on novel data. We consider the $(\lambda_z, \lambda_\varepsilon)$-combination that yields the best performance on these data to be optimal. We use this optimal penalty parameter combination to arrive at the final estimates of the two precision matrices. Without novel data unavailable for performance evaluation, they are mimicked by sample splitting. This splits the data into F equally sized groups (henceforth called splits). The splits are left-out one at the time to represent the "novel" data. Data from the remaining splits are used to obtain the precision matrices' estimates, while their performance is assessed on the "novel" data from the left-out split. Each split plays the role of "novel' data once, which results in F performance estimates. We take the average of the F performances to be indicative of the performance of the precision matrix estimators for the employed $(\lambda_z, \lambda_\varepsilon)$-combination.

The most commonly used performance measure for the selection of the penalty parameters of penalized precision matrix estimators is the cross-validation log-likelihood. That is, the averaged (over the splits) log-likelihood of data from the left-out split given the estimates derived from the data of all-but-the-left-out splits. We use this criterion here too. For practical purposes, the log-likelihood needs to be evaluated computationally efficiently, as for the cross-validated log-likelihood requires the calculation of the log-likelihood F times for each $(\lambda_z, \lambda_\varepsilon)$-combination. To achieve this efficiency, let $\mathbf{V}_{ze}\mathbf{D}_{ze}\mathbf{V}_{ze}^\top$ be the eigen-decomposition of $\mathbf{\Omega}_\varepsilon^{1/2}\mathbf{\Omega}_z^{-1}\mathbf{\Omega}_\varepsilon^{1/2}$ with the $p \times p$-dimensional matrices $\mathbf{V}_{ze}$ and diagonal $\mathbf{D}_{ze}$ that contain the eigenvectors and -values as columns and on its diagonal, respectively. We then write the log-likelihood as:

$$\log(|\mathbf{\Omega}_\varepsilon|)\sum_{i=1}^{n} K_i - \sum_{i=1}^{n}\sum_{j=1}^{p} \log[K_i(\mathbf{D}_{ze})_{jj} + 1] - \text{tr}\left[\mathbf{\Omega}_\varepsilon \sum_{i=1}^{n}(K_i - 1)\mathbf{S}_{\varepsilon,i}\right] - \text{tr}\left[\sum_{i=1}^{n} \mathbf{\Omega}_\varepsilon^{-1/2}\mathbf{V}_{ze}(K_i\mathbf{D}_{ze} + \mathbf{I}_{pp})^{-1}\mathbf{V}_{ze}^\top\mathbf{\Omega}_\varepsilon^{-1/2}\mathbf{S}_{y,i}\right],$$

with $\mathbf{S}_{y,i}$ and $\mathbf{S}_{\varepsilon,i}$ as defined at the end of Section 2. Clearly, this avoids the formation and inversion of the $K_i\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_\varepsilon^{-1}$ matrix for each different number of replicates.

A study design with replicates allows for various ways of construction the F cross-validation splits. Three strategies may be conceived:

- *Replicate-based splitting:* Form a $(\sum_{i=1}^{n} K_i) \times p$ dimensional matrix with each row containing the data from a replicate. Then, divide rows randomly over the F splits.
- *Sample-based splitting:* Divide the samples randomly over the F splits. A split's data are then formed by the replicated data of the samples that have been assigned to the split.
- *Stratified splitting:* Stratify for the number of replicates while randomly assigning the same number of samples to each split. This ensures that the distribution of the number of replicates in each split is representative of the prevalence of the K_i's encountered in the study.

Taken at face value the above splitting strategies may all seem valid. However, the first two may yield splits that are neither representative nor balanced. In particular, the first strategy may, when F is large or the K_i are small, yield splits that are unlikely to contain replicated observations of the same sample. In practice, K_i is usually small, rendering replicate-based sampling a poor choice. The sample-based strategy may, when the number of replicates is unbalanced among samples, occasionally produce splits that accidently comprise much more data than others. The resulting cross-validated performance need then not be representative. Hence, generally, the third strategy is the safest option, which is employed in the remainder. However, if the number of replicates is common to all samples, stratified and sample-based splitting are equivalent. In a simulation study we compared the consequence of the sample-based and stratified splitting for the reconstruction of the conditional independence graph, as well as the fold size. The results are presented in SM If of Appendix S1. In this study a sample's number of replicates equals either one or four, randomly chosen in a two-to-one ratio. Little to no difference is observed in the reconstruction performance. This suggests that generally both splitting strategies are viable.

The optimal $(\lambda_z, \lambda_\varepsilon)$-combination, that is, the combination of penalty parameter values that yields the precision matrix estimates with the best cross-validated performance, can—in principle—be found by a simple exhaustive grid search. Here we use the quasi-Newton approach of Byrd et al[20] available through the `optim`-function of R.[21] Alternatively, a tailor-made gradient ascent or descent approach may be developed as outlined in the work of Feng and Simon.[22] But—in light of the limited number of penalty parameters to be optimized over—the latter is not expected to give a substantial computational gain in comparison to the employed quasi-Newton approach and is therefore not pursued.

4 | SIMULATION

We study the quantification of the signal from replicated data through simulation. In the simulation, the support of employed signal precision matrices corresponds to archetypical topologies as a chain, block, and scale-free network. The error precision matrices are either diagonal or have a common nonzero off-diagonal conditional covariance. Furthermore, the sample size $n \in \{10, 25, 50, 75, 100\}$, the dimension $p \in \{10, 25, 50\}$, and the number of replicates $K_i \in \{1, 2, 3, 4\}$. Each setup is repeated a hundred times. Full simulation setup details are given in the SM Ia of Appendix S1. The aforementioned quantification comprises the performance of the signal precision matrix estimator by the Frobenius loss as well as the ability to reconstruct the signal conditional independence graph through the pAUC and AUC (partial Area Under the Curve). The latter two statistics are calculated using an edge selection procedure based on the absolute value of the partial correlations obtained from estimated signal precision matrix. We place a threshold on these absolute values, and select the edges with values exceeding the threshold. The selected edges are compared to the true graph to obtain the specificity and sensitivity. The threshold is varied over the unit interval. From the resulting (specificity, sensitivity)-pairs, we calculate the AUC and pAUC. Using the aforementioned performance measures, we first study the effect of (the number of) replicates for various sample sizes and dimensions but also parameter choices. Results are presented as Figure 1 and those in the SM Ib of Appendix S1, which—for reasons of space and brevity—are limited to one representative combination of signal and error precision matrix. The results indicate that the performance of the estimator improves in all senses specified above. This performance gain is largest from $K_i = 1$ to $K_i = 2$ and levels off for larger number of replicates. However, instead of characterizing each sample in duplicate, it is generally more rewarding to double the sample size as that appears to yield a larger improvement in performance. New samples are of course easily acquired in a simulation study but this need not necessarily be a trivial exercise in a clinical context. Finally, it should be kept in mind that these conclusion are confined to the particulars of the parameter choices. For instance, simulations (not shown) with a smaller signal-to-noise ratio reveal the levelling off is observed at larger K_i.

A tangible implication of two replicates ($K = 2$) over that of a single ($K = 1$) observation per individuals is an improvement of the estimates. The elements of the estimated precision matrix are, on average over all employed settings and topologies, 0.03 closer to their true value. Similarly, the off-diagonal elements of the corresponding partial correlation matrices are, again on the same average, 0.02 closer to their true value. This improvement is largest for the larger sample sizes, the smaller dimensions, and the larger elements of $\mathbf{\Omega}_z$. It can then go up 0.1 for off-diagonal elements, and even over 0.2 for diagonal elements (of the precision matrix).

Another takeaway of this simulation can be deduced from the scale of the y-axis of Figure 1 and its companions in the Appendix S1. On the basis of pure chance, one would expect the pAUCs to be around 0.005. Simultaneously, the maximum achievable pAUC is 0.1. While the results clearly exceed the chance benchmark, they are not close to their maximum. This demonstrates the notorious difficulty of the network reconstruction problem for moderate dimensions ($p = 50$). The difficulty is readily grasped when realizing that, for a p variate, one needs to estimate $\frac{1}{2}p(p + 1)$ parameters (counting on those related to the signal) from a small number of samples. However, the achieved pAUCs are also due to the chosen simulation settings. Other settings, for example, less noise or stronger effect sizes, would have yielded a better pAUC. But, for instance, we based the sample size range on practice, where our in-house studies rarely exceed a hundred samples and often involve fewer. Nonetheless, the pAUC plots—as the simulation intends—clearly show the effect of the inclusion of replicates, in particular in relation to the sample size and dimension. Finally, the reported pAUCs serve as a warning that, for a small sample size and few duplicates, results are in urgent need of validation.

Secondly, we assess whether the full design needs replication, or that it is best to replicate only part of them. Hereto we adopt the settings of the previous simulation, with the following modification. We set the total number of measurement $\sum_{i=1}^{n} K_i$, with $K_i \in \{1, 2\}$ for all i, equal to hundred. Under this restriction, we vary the number of samples with a single observation and two replicates. With this study design, the above simulation is repeated. The bottom left panel of Figure 1 shows the achieved pAUC against the number of replicated samples. At first, there is a clear gain with each additional replicated sample, although not so obvious for the $p = 50$ case. This gain, however, levels off after a certain number—the precise number depends among others on the dimension and signal-to-noise level—of replicated samples. It even goes down when further samples are replicated. This indicates that at some point it is more worthy to include

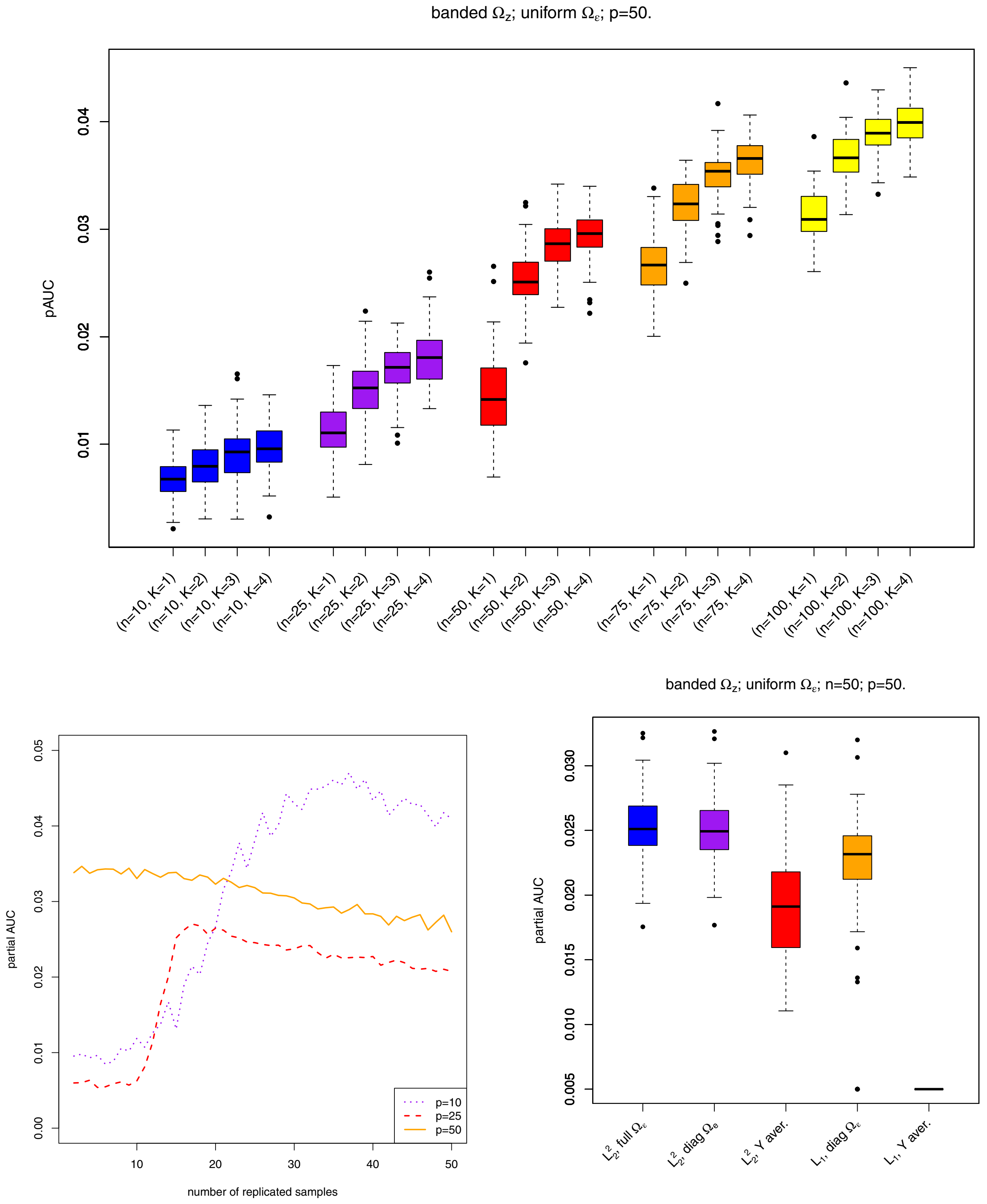

FIGURE 1 Various simulation results w.r.t. edge recovery for a banded signal precision matrix $\mathbf{\Omega}_z$ and a uniform error precision matrix $\mathbf{\Omega}_\varepsilon$. All plots show the partial AUC, integrated w.r.t. 1 – specificity from 0 to 0.1, of 100 simulation runs. In the top panel $p = 50$ and the pAUCs are plotted against various (n, K_i)-combinations. The left bottom panel plots, for $p = 10, 25, 50$, the averaged pAUC vs the number of replicated samples with $\sum_{i=1}^{n} K_i = 100$ and all $K_i \in \{1, 2\}$. The right bottom panel, in which $(n, p, K) = (50, 50, 2)$, shows boxplots of pAUCs of five methods. Legend for the labels at its tick marks: 'L_2^2, full $\mathbf{\Omega}_\varepsilon$': Ridge penalized EM algorithm without the diagonal error precision matrix assumption; "L_2^2, diag $\mathbf{\Omega}_\varepsilon$": Ridge penalized EM algorithm with the diagonal error precision matrix assumption; "L_2^2, Y average": Ridge penalized estimation of $\mathbf{\Omega}_y$ from replicate-wise averaged data (ie, $\{K_i^{-1} \sum_{k_i=1}^{K_i} \mathbf{Y}_{i,k_i}\}_{i=1}^{n}$; "$L_1$, diag $\mathbf{\Omega}_\varepsilon$": Lasso penalized EM algorithm with the diagonal error precision matrix assumption; "L_1, Y average": Lasso penalized estimation of $\mathbf{\Omega}_y$ from replicate-wise averaged data (ie, $\{K_i^{-1} \sum_{k_i=1}^{K_i} \mathbf{Y}_{i,k_i}\}_{i=1}^{n}$ [Colour figure can be viewed at wileyonlinelibrary.com]

biological—should they be available—rather than technical replicates. Especially, when $p=50$ the gain of replication is limited and more biological samples are to be preferred.

Thirdly, we compare in simulation the proposed method to some obvious competitors: (i) the lasso penalized EM algorithm with a diagonal error precision matrix, (ii) the ridge precision estimator[5] with replicate-wise averaged data, and (iii) the graphical lasso precision estimator[4] with replicate-wise average data. The simulation setup is as above but with $K_i=2$ throughout and $n \in \{10, 25, 50\}$. Results are presented as boxplots in the figures of SM Ie of Appendix S1, again limited to a representative combination of signal and error precision matrix. The main takeaways are two-fold. Firstly, the ridge penalized methods generally outperform their lasso counterparts, in particular for the larger p, in terms of network reconstruction. Secondly, network reconstruction by means of the ridge and lasso precision estimators from averaged data works reasonably well (the latter only with large n and small p). That is, the support of the estimated network based on averaged data do not substantially differ with respect to the AUC-type measures. However, the values of the estimated signal precision matrices on the basis of averaged data can differ substantially in terms of the Frobenius norm. A more detailed conclusion is given in SM Ie of Appendix S1. Originally, we included a lasso penalized EM algorithm with a penalty on both precision matrices. The resulting algorithm's convergence was slow, while the search for optimal cross-validated penalty parameters was prohibitively slow. Moreover, the results of the other lasso penalized methods, that is, (i) and (iii), suggest its performance will not exceed that of their ridge counterparts.

5 | ILLUSTRATION

We present an illustration of the use of the presented methodology through a re-analysis of several oncogenomics studies with replicated observations. The aim of this re-analysis is multifold: (i) to clarify the consequences of the conditional independence graph reconstruction from an error-diluted signal, (ii) to assess the tenability of the independence assumption among the errors as implied by a diagonal $\mathbf{\Omega}_\epsilon$ for the current purpose, and (iii) to elucidate the differences between conditional independence graphs reconstructed from replicated and nonreplicated data.

The data stem from three TCGA (The Cancer Genome Atlas) studies[23-25] into the molecular characterization of the cancer of three tissue types, breast ($n=526$), lung ($n=151$), and ovary ($n=294$). Each study interrogated a sample's transcriptome twice (ie, $K_i=2$ for all i), by both gene expression arrays and RNA sequencing. These data have been downloaded using the `TCGA2STAT`-package.[26] Subsequently, each dataset has been subsetted into ten smaller ones, each formed by restricting the original dataset to a subset of the genes. The preserved genes in each subsetted dataset map to one of ten signaling pathways that are believed to be involved in cancer. The definitions of these pathways are taken from KEGG[27] and available in `R` through the `KEGG.db`-package[28] as a set of so-called Entrez-identifiers. These identifiers are matched to those of the genes present in the datasets. The latter step required conversion of the gene names to their Entrez-identifiers for which we have used the `biomaRt`-package.[29] The pathways' names and their dataset-specific dimension (ranging from $p=29$ to $p=247$) and sample size are tabulated in SM IIa of Appendix S1. Finally, to meet the distributional assumptions of the presented model the data have been Gaussianized variate-wise, an operation that preserves the conditional independencies among variates.[30] Other assumptions are checked visually (see SM IIc of Appendix S1), and found to be unproblematic.

We analyze the data in the following ways. We fit the presented model, with and without the diagonal assumption on the error precision matrix, with the penalty parameter(s) chosen through stratified 10-fold cross-validation. Additionally, we learn the platform-specific, that is, array and sequencing, precision matrices from the data using the ridge precision matrix estimator[5] that uses a penalty parameter found through 10-fold cross-validation. In the remainder of this section we scrutinize the resulting precision matrices to meet the aims formulated at the beginning of this illustration.

5.1 | The effect of the error

The deconvolution of signal and error by fitting the "signal+error" model facilitates the study of the consequence of the error on the learning of the conditional independence graph. This study comprises (i) the quantification of the contribution of signal and error to the observation, (ii) the comparison of partial correlations derived from the signal and error-diluted observation precision matrices, and (iii) the therefrom inferred conditional independence graphs.

The fitted model enables us to investigate whether the observations are dominated by either the signal or the error. Hereto we employ the mutual information, a generalized correlation measure, that measures the dilution of the signal by the addition of the error (or vice versa). Concentrating on the former, the mutual information between $\mathbf{Y}_{i,k}$ and $\mathbf{Z}_i$ is $\mathcal{I}(\mathbf{Y}_{i,k};\mathbf{Z}_i) = \mathcal{H}(\mathbf{Y}_{i,k}) - \mathcal{H}(\mathbf{Y}_{i,k} \,|\, \mathbf{Z}_i)$, where, for example, $\mathcal{H}(\mathbf{Y}_{i,k})$ is the (differential) entropy of $\mathbf{Y}_{i,k}$. Large values of $\mathcal{I}(\mathbf{Y}_{i,k};\mathbf{Z}_i)$ indicate that $\mathbf{Z}_i$ contains a lot information on $\mathbf{Y}_{i,k}$, whereas $\mathcal{I}(\mathbf{Y}_{i,k};\mathbf{Z}_i) = 0$ means the random variables are independent. Here, in the multivariate normal case, by theorem 9.4.1 of Clover and Thomas,[31] $\mathcal{I}(\mathbf{Y}_{i,k};\mathbf{Z}_i) = \frac{1}{2}\log(|\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_\varepsilon^{-1}|/|\mathbf{\Omega}_\varepsilon^{-1}|)$. Similarly, $\mathcal{I}(\mathbf{Y}_{i,k};\boldsymbol{\varepsilon}_{i,k}) = \frac{1}{2}\log(|\mathbf{\Omega}_z^{-1} + \mathbf{\Omega}_\varepsilon^{-1}|/|\mathbf{\Omega}_z^{-1}|)$. For each (dataset, pathway)-combination we evaluate these mutual informations by plug-in estimates of the precision matrices. These results are tabulated in SM IIe of Appendix S1. These tables show that, structurally over all (dataset, pathway)-combinations, the $\mathcal{I}(\mathbf{Y}_{i,k};\mathbf{Z}_i)$ are substantially larger than $\mathcal{I}(\mathbf{Y}_{i,k};\boldsymbol{\varepsilon}_{i,k})$. From this we conclude that the observations are dominated by the signal and not the error. This conclusion is corroborated by exploratory analyses presented in SM IIe of Appendix S1. Consequently, when replicates are not available, the inference of the signal-related conditional independence graph directly from the estimated observation-related precision matrix $\mathbf{\Omega}_y$ is not completely in vain.

We compare the distributions of the partial correlations, the basis of the inference of the conditional independence graph, derived from the estimated signal and observation precision matrices, $\widehat{\mathbf{\Omega}}_z$ and $\widehat{\mathbf{\Omega}}_y = [(\widehat{\mathbf{\Omega}}_z)^{-1} + (\widehat{\mathbf{\Omega}}_\varepsilon)^{-1}]^{-1}$. Hereto we generate (i) qq-plots (see SM IIf of Appendix S1) and (ii) the densities (not shown) of the differences between corresponding partial correlations. The qq-plots suggest that both partial correlation distributions are reasonably similar, but with differences appearing mainly in the tails. The densities of the partial correlations differences confirm this, as most mass is concentrated around and close to zero. However, the different tail behavior implies that the error indeed obscures true edges as well as introduces spurious ones in the inferred conditional independence graph.

We now quantify the effect of the error on the inferred conditional independence graph as follows. This graph is inferred from both partial correlation matrices, that is, the ones derived from the signal and observation precision matrix estimates $\widehat{\mathbf{\Omega}}_z$ and $\widehat{\mathbf{\Omega}}_y = [(\widehat{\mathbf{\Omega}}_z)^{-1} + (\widehat{\mathbf{\Omega}}_\varepsilon)^{-1}]^{-1}$. The graph is formed by simply taking the top r, $r = 1, \ldots, 250$, largest (in an absolute sense) unique partial correlations from both matrices. The percentage of overlapping edges among the selected edges between the two graphs is plotted against the number of selected edges (see Figure 2), again for each (dataset, pathway)-combination. Expectedly, this percentage is unstable for small r, but settles for larger ones. On average, over data sets and pathways, it settles around approximately 70%. Would we translate this to the inference of molecular networks through the learning of conditional independence graphs from a single platform, it suggests that little over one in four absent/present edges reported in the literature is either a false positive or false negative.

Finally, we illustrate the diluting effect of the error on the estimation partial correlations. Hereto the signals $\mathbf{Z}_i$ for $i = 1, \ldots, n$ are estimated by $\mathbb{E}(\mathbf{Z}_i \,|\, \mathbf{Y}_{i,\text{array}}, \mathbf{Y}_{i,\text{seq}}, \widehat{\mathbf{\Omega}}_z, \widehat{\mathbf{\Omega}}_\varepsilon)$ with data and estimates stemming from the apoptosis pathway of the TCGA lung study. We then simulate observed data by $\mathbf{Y}_i = \widehat{\mathbf{Z}}_i + \boldsymbol{\varepsilon}_i$ with $\boldsymbol{\varepsilon}_i$ drawn from $\mathcal{N}(\mathbf{0}_p, \widehat{\mathbf{\Omega}}_\varepsilon^{-1})$. The samples are thus unreplicated. The signal and 'observed' partial correlations are then obtained from the standardized inverse of their sample covariance matrices $\frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i\mathbf{Z}_i^{\top}$ and $\frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_i\mathbf{Y}_i^{\top}$, respectively. To capture the spread in the latter it is evaluated for a hundred error draws. Boxplots of the resulting hundred "observation" partial correlations corresponding to thirty randomly selected edges are displayed in Figure 2. Their "signal" partial correlations are plotted on top of them as blue diamonds. This reveals that indeed some partial correlations are clearly weakened by dilution of the signal by the error. Simultaneously, others are strengthened possibly introducing spuriously inferred edges. The partial correlations estimated from an unreplicated error diluted signal can thus differ substantially from those obtained from the signal itself. This bears consequences on the reconstruction of the network. To illustrate one of the implications we infer the network from the top 100 strongest (in an absolute sense) partial corrections derived from the standardized $\widehat{\mathbf{\Omega}}_z$. This yields a network of 13 unconnected nodes and one large connected component involving 66 nodes. The large connected component is the topological feature of interest. We assess whether it persists when the network is reconstructed from an error diluted signal. Such a signal is created as above, from which the corresponding partial correlation matrix is estimated, and in turn a network reconstructed by selection of its top 100 strongest edges. This exercise is repeated a hundred times. The hundred networks derived from the error diluted signal all exhibit a large connected component. In over 85% of these networks the size of this component involves 50 or more nodes. Hence, without replicated samples the prominent network feature is generally preserved, but it is also partially obscured due to error dilution.

5.2 | The diagonal $\mathbf{\Omega}_\varepsilon$ assumption

The assumption of a diagonal $\mathbf{\Omega}_\varepsilon$ discussed in Section 3.1 is evaluated. Previously, we proposed the assumption for computational reasons, in particular when the penalty parameter is chosen via cross-validation. Here we study its effect on the reconstruction of the conditional independence graph in real-data.

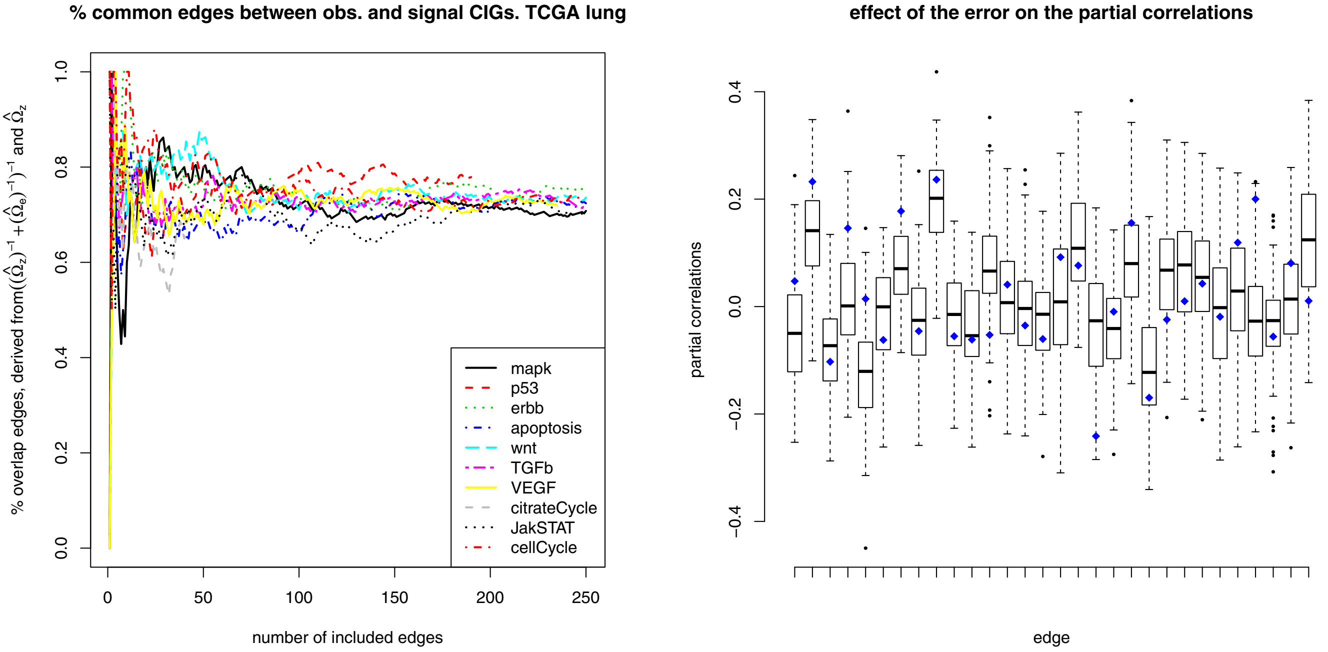

FIGURE 2 Left panel: the percentage of overlapping edges (y-axis) between the conditional independence graphs formed by selecting the top r (x-axis) strongest (in an absolute sense) partial correlations from the standardized signal precision matrix $\hat{\mathbf{\Omega}}_z$ and the "observation" precision matrix $\hat{\mathbf{\Omega}}_y = [(\hat{\mathbf{\Omega}}_z)^{-1} + (\hat{\mathbf{\Omega}}_\varepsilon)^{-1}]^{-1}$. Each line represents a different pathway and connects the percentages of overlapping edges found for a top of varying sizes r, $r = 1, \dots, 250$. Right panel: boxplots of partial correlations of randomly selected edges evaluated from a fixed signal $\mathbf{Z}_i$ diluted with varying errors $\boldsymbol{\varepsilon}_i$. For reference the partial correlations from the undiluted signals are added as blue diamonds [Colour figure can be viewed at wileyonlinelibrary.com]

For starters we compare the models with a full and diagonal $\mathbf{\Omega}_\varepsilon$ by means of the Aikake's Information Criterion (AIC). The AIC balances the model's fit with its parsimony. For the model with a full $\mathbf{\Omega}_\varepsilon$ the AIC is:

$$AIC = 2p(p+1) - 2\sum_{i=1}^{n} \log[P(\mathbf{Y}_{i,1}, \dots, \mathbf{Y}_{i,K_i}; \hat{\mathbf{\Omega}}_z, \hat{\mathbf{\Omega}}_\varepsilon)],$$

that is, twice the number of model parameter minus twice the log-likelihood evaluated at the estimated model parameters under the full $\mathbf{\Omega}_\varepsilon$ assumption. For the model with a diagonal $\mathbf{\Omega}_\varepsilon$ the first summand on the right-hand side of the preceding display is replaced by $p(p+1) + 2p$ and the corresponding estimators are used in the log-likelihood. These estimated AICs are reported in SM IId of Appendix S1. They reveal that the AICs of the model with a full $\mathbf{\Omega}_\varepsilon$ are better (ie, smaller) than the model with a diagonal $\mathbf{\Omega}_\varepsilon$. Hence, the improvement of the description of the data by the more elaborated model over the simpler one outweighs the use of $\frac{1}{2}p(p-1)$ additional parameters by the former. This suggests the full model is to be preferred when used for the reconstruction of the conditional independence graph.

The model with a diagonal $\mathbf{\Omega}_\varepsilon$ may not be preferred on the basis of the AIC, it could still be a good basis for the reconstruction of the conditional independence graph. As in Section 5.1 qq-plots of the partial correlations, derived from the $\mathbf{\Omega}_z$ estimate under both error assumptions, are drawn for every (data set, pathway)-combination (see SM IIf of Appendix S1). These plots reveal little difference in distribution of these partial correlations. Additionally, the densities of the differences of corresponding partial correlations of both models are plotted (not shown). Generally, these densities are tightly concentrated around zero, suggesting the estimate of $\mathbf{\Omega}_z$ under the assumption of a diagonal $\mathbf{\Omega}_\varepsilon$ may still be a good basis for the reconstruction of the conditional independence graph. This is quantified, as in Section 5.1, by the percentage of overlapping edges between the conditional independence graphs reconstructed from both $\mathbf{\Omega}_z$ estimates, and selecting only the top r, $r = 1, \dots, 250$, largest (in an absolute sense) partial correlations. For each (dataset, pathway)-combination we plot these percentages against the number of selected edges r. In all cases the percentage of overlapping edges between the two reconstructed networks exceeds the 85% and is on average around 90%. Hence, for initial screening purposes a simpler model may suffice, but the computational efficiency gain comes at a cost.

5.3 | The platform differences

In the spirit of the MAQC we compare the reconstruction of the CIGs from individual—but also joint—platform data (all plots are deferred to SM IIf of Appendix S1). In the remainder we refer to these graphs as the "microarray CIG," the "RNA-seq CIG" and the "joint CIG." The percentage of overlap among the top r edges of the micro-array and RNA-seq CIGs varies roughly between 35% and 55% (cf, the SM IIf of Appendix S1) over pathways and datasets. This suggests that roughly only between a third and a half of the edges reported in the literature will reproduce in subsequent studies when using a different platform.

In our comparison of the individual platforms' CIG to the joint one, we assume that (a) the variation in the data only comprises sampling variation and that due to the use of the two different platforms and (b) these variation components can be estimated adequately and without (!) too much error by the proposed penalized EM algorithm in combination with a cross-validated penalty parameter. The percentage of overlap in the top r edges of the joint and either the micro-array or RNA-seq CIGs fluctuates around approximately 60% and 65%, respectively, over pathways and datasets. The outlying percentages for the MAPK pathway in the ovarian data are due to extremely large cross-validated penalties in both platform specific precision matrix estimates. The overlap of the "joint CIG" with the RNA-seq one is systematically a little larger with that of the micro-array platform. Irrespectively of this minor difference, these percentages suggest that—although unknown which—65% of the gene-gene interactions reported in the literature are correctly identified, should the aforementioned assumptions be tenable. This 65% is slightly smaller but otherwise in line with the approximately 70% found in Section 5.1, when investigating the effect of the error. The former percentage can be dissected into the overlap percentages among the top r edges between:

- the joint CIG and the intersection of the microarray and RNA-seq CIGs. This overlap percentage ranges from 35% to 55%, depending on pathway and data set. In particular, the plots also indicate that, if an edge is in the overlap of the platform specific CIGs, it is most likely to be in the joint CIG.
- the joint and microarray CIGs that are not present in the RNA-seq CIG. This ranges more or less from 15% to 20%, while vice versa the joint and RNA-seq CIGs that are not present in the micro-array CIG fluctuate between 20% and 25%. The latter's larger overlap is in line with the overall larger overlap between these two CIGs. Irrespectively, this indicates that indeed there are platform specific edges.
- the joint CIG that are not present in either the microarray or RNA-seq CIGs. It ranges from 5% to 15% (and sometimes as high as 25%) over the pathways and datasets. This reveals the amount of obscured edges by use of a particular platform.

These percentages should be related to the probability of an edge common to two independently reconstructed networks with the same number of nodes p and an equal number of selected edges r. For $p = 50$ and $r = 250$ is approximately 4.16% (and lower for smaller r or larger p). Hence, as the observed percentage easily exceeds this reference percentage of 4.16%, there is definitely shared information between the platform-specific CIGs. Although not perfect, it represents the cohesion of the pathways' gene expression data.

Finally, we draw up a more specific inventory of the overlap between the joint, microarry and RNA-seq CIGs. Hereto we identify, using the TCGA lung cancer data, for all pathways the 100 strongest edges from the corresponding partial correlation matrices. For each pathway we evaluate the overlap between the all combinations of the resulting CIGs, shown in Figure 3. Would all three CIGS be identical, a bar is solid brown and reaches up to 100 on the x-axis. Similarly, without overlap among the CIGs, a bar comprises of three equally sized blocks, coloured red, yellow and blue, while reaching up to 300. In Figure 3 the bars reach—on average—to approximately 175, which amounts to a reasonable amount of overlap. Unsurprisingly, the joint CIGs share most with both other CIGs, individually and with their intersection. However, there are also approximately 15 edges present only in the joint CIG, which—if correct—are missed without replication. On a similar note, a much larger number of edges is specific to either the RNA-seq or the micro-array CIGs. Hence, using a single platform without replication, one clearly identifies a substantial amount of edges that are unlikely to reproduce.

6 | CONCLUSION

Assuming a simple "signal+noise" model we showed in Section 2 the possible consequences of ignoring variation due to other sources than sampling for the reconstruction of the cohesion among the variates of a Gaussian random variable:

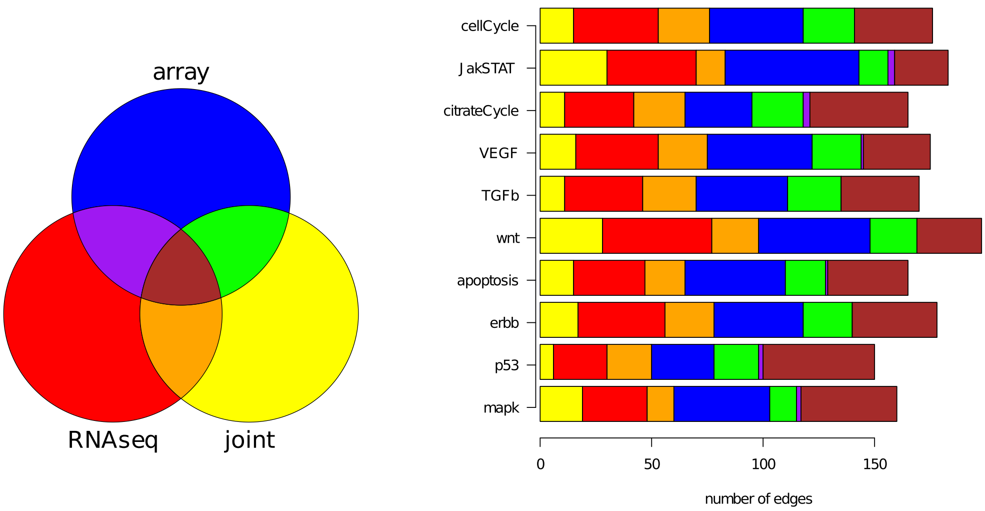

FIGURE 3 On the right, a horizontal bar plot of—per pathway—the number of overlapping edges between the top 100 strongest edges of the CIGs reconstructed from the RNA-seq, micro-array and the joint data. The left panel represents the accompanying color legend via a venn diagram [Colour figure can be viewed at wileyonlinelibrary.com]

conditional dependencies may be obscured and spurious ones introduced. We pointed out that this may be overcome when observations have been replicated as different sources of variation can be separated. We presented methodology for the estimation of the parameters associated with these sources and that harbor the sought conditional (in)dependencies. Simulations showed that most is gained from duplication but that, for example, triplicated observations add little. It also revealed that, when pragmatically using replicate-wise averaged data instead of the more complicated proposed "signal+noise" model-based approach, the support of the signal precision matrix can be reconstructed quite well but the corresponding estimated values of the precision matrix can be inaccurate. Finally, through an extensive re-analysis of data from oncogenomics studies with replicated observations the effect of omission of replicates but also the gain of their inclusion has been tangibly illustrated. In particular, it provides insight in the reproducibility of published gene-gene interaction networks, which indicates that care is to be taken with the validity of these networks.

A further note of caution is needed. Sofar false-positive and -negative edges of the reconstructed conditional independence graph have only been attributed to the presence of the variation introduced by the use of different platforms. On one hand, this ignores the uncertainty in the estimation due to the use of a sample of finite size that introduces falsely inferred absent and present edges. On the other, the focus is—due to the design of the employed TCGA studies—on the error quantifiable from technical replicates. This ignores the fact that mRNA levels may vary considerably over the day. This biological within-sample variation cannot be quantified from the used TCGA studies. That would require studies with a longitudinal setup in which samples are characterized at multiple instances. And for its analysis different statistical methodology is needed. Both are the subject of follow-up research.

ACKNOWLEDGEMENTS

This project has received funding from the Euratom research and training programme 2014–2018 under grant agreement No 755523.

DATA AVAILABILITY

The data that support the findings of this study are openly available in at the TCGA portal, here accessed through the R-package TCGA2STAT. The presented methodology has been implemented in the R statistical computing language[21] and is incorporated in the porridge-package[32] available via CRAN repository (https://cran.r-project.org/web/packages/porridge).

ORCID

Wessel N. van Wieringen https://orcid.org/0000-0002-5100-9123

REFERENCES

1. Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. *J Multivar Anal*. 2004;90(1):196-212.
2. Whittaker J. *Graphical Models in Applied Multivariate Statistics*. New York, NY: John Wiley; 1990.
3. Lauritzen SL. *Graphical Models*. Oxford, UK: Oxford University Press; 1996.

4. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008;9(3):432-441.
5. van Wieringen WN, Peeters CFW. Ridge estimation of the inverse covariance matrix from high-dimensional data. *Comput Stat Data Anal*. 2016;103:284-303.
6. Shewhart WA. *Economic Control of Quality of Manufactured Product*. London: Macmillan And Co Ltd; 1931.
7. Wainwright MJ. *High-Dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge, MA: Cambridge University Press; 2019.
8. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*. 2006;24(9):1151-1161.
9. MAQC Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol*. 2010;28(8):827-838.
10. Zhang W, Yu Y, Falk Hertwig F, et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol*. 2015;16:133.
11. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? *PLoS Comput Biol*. 2011;7(1):e1001057.
12. Bellot P, Olsen C, Salembier P, Oliveras-Vergés A, Meyer PE. NetBenchmark: a bioconductor package for reproducible benchmarks of gene regulatory network inference. *BMC Bioinform*. 2015;16(1):312.
13. Vinciotti V, Wit EC, Jansen R, et al. Consistency of biological networks inferred from microarray and sequencing data. *BMC Bioinform*. 2016;17(1):254.
14. Miller KS. On the inverse of the sum of matrices. *Math Mag*. 1981;54(2):67-72.
15. Riebler A, Held L, Rue H. Estimation and extrapolation of time trends in registry data-borrowing strength from related populations. *Ann Appl Stat*. 2012;6(1):304-333.
16. Harville DA. *Matrix Algebra from a Statistician's Perspective*. New York, NY: Springer; 2008.
17. Titterington DM, Smith AF, Makov UE. *Statistical Analysis of Finite Mixture Distributions*. New York, NY: John Wiley; 1985.
18. Anderson TW. *An Introduction to Multivariate Statistical Analysis*. 3rd ed. New York, NY: John Wiley; 2003.
19. Zhu Z, Melnykov V. ManlyMix: an R package for manly mixture modeling. *R J*. 2017;9(2):176-197.
20. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput*. 1995;16:1190-1208.
21. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2018 https://www.R-project.org/.
22. Feng J, Simon N. Gradient-based regularization parameter selection for problems with nonsmooth penalty functions. *J Comput Graph Stat*. 2018;27(2):426-435.
23. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474:609-615.
24. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489:519-525.
25. Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490:61-70.
26. Wan YW, Allen GI, Anderson ML, Liu Z. TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. R package version 1.2; 2015. https://CRAN.R-project.org/package=TCGA2STAT.
27. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucl Acids Res*. 1999;27(1):28-34.
28. Carlson M. KEGG.db: a set of annotation maps for KEGG. R package version 3.2.3; 2016. https://www.bioconductor.org/packages/release/data/annotation/html/KEGG.db.html.
29. Durinck S, Spellman PT, Birney W, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. R package version 2.34.2. *Nat Protoc*. 2009;4:1184-1191. https://bioconductor.org/packages/release/bioc/html/biomaRt.html.
30. Liu H, Lafferty J, Wasserman L. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J Mach Learn Res*. 2009;10(Oct):2295-2328.
31. Clover TM, Thomas JA. *Elements of Information Theory*. New York, NY: John Wiley; 2006.
32. van Wieringen WN. Porridge: ridge-type estimation of a potpourri of models. R package version 0.01; 2019. https://CRAN.R-project.org/package=porridge.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: van Wieringen WN, Chen Y. Penalized estimation of the Gaussian graphical model from data with replicates. *Statistics in Medicine*. 2021;40:4279–4293. https://doi.org/10.1002/sim.9028