Computational Statistics and Data Analysis 156 (2021) 107085

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian regularization of Gaussian graphical models with measurement error

Michael Byrd [a,*], Linh H. Nghiem [b], Monnie McGee [a]

[a] Department of Statistical Science, Southern Methodist University, TX, 75206, USA
[b] Research School of Finance, Actuarial Studies, and Statistics, Australian National University, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 31 March 2020
Received in revised form 25 August 2020
Accepted 26 August 2020
Available online 3 September 2020

Keywords:
Data contamination
Regularization
Gene networks
Graphical models

ABSTRACT

A framework for determining and estimating the conditional pairwise relationships of variables in high dimensional settings when the observed samples are contaminated with measurement error is proposed. The framework is motivated by the task of establishing gene regulatory networks from microarray studies, in which measurements are taken for a large number of genes from a small sample size, but often measured imperfectly. When no measurement error is present, this problem is often solved by estimating the precision matrix under sparsity constraints. However, when measurement error is present, not correcting for it leads to inconsistent estimates of the precision matrix and poor identification of relationships. To this end, a recent iterative imputation technique developed in the context of missing data is utilized to correct for the biases in the estimates imposed from the contamination. This technique is showcased with a recent variant of the spike-and-slab Lasso to obtain a point estimate of the precision matrix. Simulation studies show that the new method outperforms the naïve method that ignores measurement error in both identification and estimation accuracy. The new method is applied to establish a conditional gene network from a microarray dataset.

1. Introduction

In biomedical settings, it is often of interest to use gene expression microarray data to construct a gene regulatory network for metabolic processes (Segal et al., 2003). In such studies, the gene measurements follow a multivariate Gaussian distribution, where the inverse covariance matrix, known as the precision matrix, characterizes conditional dependence between two genes (called two dimensions, or two features). This is accomplished by noting that if an element of the precision matrix is 0, then the two dimensions are conditionally independent; see Lauritzen (1996) for a review. This setting, often referred to as a Gaussian graphical model, is where our analysis takes place.

Estimating the precision matrix is a difficult task when the number of observations n is often much less than the dimension of the features d. A naïve approach is to estimate the precision matrix by the inverse of the empirical covariance matrix; this estimate, however, is known to perform poorly and is ill-posed when $n < d$ (Johnstone et al., 2001). The common approach is to assume that the precision matrix is sparse (Dempster, 1972); that is, we assume the precision matrix's off-diagonal elements are mostly 0. As a result, most pairs of variables are conditionally independent. The sparsity assumption has led to different lines of research with regularized models to estimate the precision matrix. While some approaches utilize a sparse regression technique that estimates the precision by iteratively regressing each variable on

* Correspondence to: 3225 Daniel, Suite 144 Heroy Building Dallas TX 75205, USA.
E-mail addresses: mcbyrd@smu.edu (M. Byrd), linh.nghiem@anu.edu.au (L.H. Nghiem), mmcgee@smu.edu (M. McGee).

https://doi.org/10.1016/j.csda.2020.107085
0167-9473/

the remaining variables, for instance Khare et al. (2015), we instead focus on the direct likelihood approach. The direct likelihood approach optimizes the full likelihood function with an element-wise penalty on the precision matrix; common examples being graphical lasso (Friedman et al., 2008), CLIME (Cai et al., 2011), and TIGER (Liu et al., 2017). We utilize a recent Bayesian optimization procedure, called BAGUS, that relies on optimization performed by the EM-algorithm, which was shown to have desirable theoretical properties, including consistent estimation of the precision matrix and selection consistency of the conditional pair-wise relationships (Gan et al., 2018).

There are many practical issues associated with Gaussian graphical models, such as hyperparameter tuning (Yuan and Lin, 2007), missing data (Liang et al., 2018), and repeated trials (Tan et al., 2016), which practitioners need to adjust for a successful analysis. We address another practical issue that is often involved with the microarray studies, measurement error. In fact, microarray studies tend to have noisy measurements because of technical variations resulting from sources such as sample preparation, labeling, and hybridization (Zakharkin et al., 2005). In other words, the observations are expression values that have been additionally perturbed with noise from some measurement process. The effects of measurement error on statistical models have been studied extensively for classical settings such as density deconvolution and regression (Carroll et al., 2006), but, to our knowledge, has not yet been well studied in the context of Gaussian graphical models, especially in high dimensional settings.

We propose a Bayesian estimator to correct for measurement error in estimating a sparse precision matrix; our new method extends the optimization procedure of Gan et al. (2018), referred to as BAGUS. While directly incorporating the estimate of the uncontaminated variable is possible, we find the incorporation of the imputation–regularization technique of Liang et al. (2018) to provide more desirable results. Our procedure imputes the mismeasured random variables, then performs BAGUS on this imputation; these steps are iterated for a small number of cycles, requiring more computation but giving better results than the naïve estimator. We prove consistency of the estimated precision matrix with the imputed procedure, and illustrate the performance in a simulation study. Finally, we apply the methodology to a microarray dataset that contains gene expression measurements of favorable histology Wilms tumor (Huang et al., 2009).

2. Contaminated Gaussian graphical models

Given a d-dimensional random vector, $\boldsymbol{x} = \{x^1, \ldots, x^d\}$, we are interested in the conditional dependence of two variables x^i and x^j, for any pair (i, j) with $1 \leq i < j \leq d$, given all the remaining variables. This conditional dependence structure is usually represented by an undirected graph $G = (V, E)$, where $V = \{1, \ldots, d\}$ is the set of nodes and $E \subseteq V \times V = \{(1, 1), (1, 2), \ldots, (d, d)\}$ is the set of edges (Lauritzen, 1996). In this representation, the two variables x^i and x^j are conditionally independent if there is no edge between node i and node j.

If the vector $\boldsymbol{x}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_x$, $\boldsymbol{x} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}_x)$, every edge corresponds to a non-zero entry in the precision matrix $\boldsymbol{\Omega}_x = \boldsymbol{\Sigma}_x^{-1}$, see Lauritzen (1996). The model in this scenario is often known as a Gaussian graphical model. In the high dimensional setting, the set of edges are usually assumed to be sparse, meaning that only a few pairs (x^i, x^j) are conditionally dependent. In the Gaussian case, this assumption implies only a few off-diagonal entries of $\boldsymbol{\Omega}_x$ are non-zero.

Suppose the data consist of *iid* observations $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$, where $\boldsymbol{w}_i = \boldsymbol{x}_i + \boldsymbol{u}_i,\ i = 1, \ldots, n$ with $\boldsymbol{x}_i \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}_x)$ and $\boldsymbol{u}_i \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}_u)$. Here, $\boldsymbol{w}_i = (w_i^1, \ldots, w_i^d)$, with the subscript and superscript denoting the observation and components respectively. Denote $\boldsymbol{W}$ as the $n \times d$ matrix of observed data. The model is equivalent to the following hierarchical representation. First, the latent random variables $\boldsymbol{x}_i$ are generated from a $N_d(\mathbf{0}, \boldsymbol{\Sigma}_x)$ distribution, and when conditioned on $\boldsymbol{x}_i$ and $\boldsymbol{\Sigma}_u$, we have $\boldsymbol{w}_i | \boldsymbol{x}_i, \boldsymbol{\Sigma}_u \sim N_d(\boldsymbol{x}_i, \boldsymbol{\Sigma}_u)$ for each $i = 1, \ldots, n$. This forms an intuitive generative process, where first $\boldsymbol{x}$ is realized, then contaminated by measurement error $\boldsymbol{u}$, and observed finally as $\boldsymbol{w}$. The problem of interest is to estimate the precision matrix $\boldsymbol{\Omega}_x$ in the high dimensional setting $n < d$.

Consider an additive measurement error model where $\boldsymbol{w} = \boldsymbol{x} + \boldsymbol{u}$ and $\boldsymbol{w}$ is the observed data. Denote $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)^T$ as measurement errors that are independent from data $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$. For $i = 1, \ldots, n$, the amount of measurement error is drawn from another multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{u}}$, $\boldsymbol{u}_i \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})$. We assume $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ to be diagonal, and hence the amount of measurement error on each variable is uncorrelated. We also assume that $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ is known or estimable from ancillary data, such as replicate measurements. This is a common occurrence, such as in microarray studies where multiple replicates of one subject are collected (Nghiem and Potgieter, 2018), and we illustrate the estimation procedure in Section 5. The contaminated variables $\boldsymbol{w}$ in general have a different conditional dependence structure from that of $\boldsymbol{x}$. Indeed, the covariance and precision matrix of $\boldsymbol{w}$ is given by

$$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_u$$

and

$$\boldsymbol{\Omega}_w = \boldsymbol{\Sigma}_w^{-1} = (\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_u)^{-1} = \boldsymbol{\Omega}_x - \boldsymbol{\Omega}_x(\boldsymbol{I} + \boldsymbol{\Sigma}_u\boldsymbol{\Omega}_x)^{-1}\boldsymbol{\Sigma}_u\boldsymbol{\Omega}_x, \tag{1}$$

respectively; here, $\boldsymbol{I}$ denotes the $d \times d$ identity matrix. Eq. (1) follows from the Kailath variant formula in Petersen et al. (2008). Furthermore, Eq. (1) suggests that $\boldsymbol{\Omega}_w$ and $\boldsymbol{\Omega}_x$ are equal if the product $\boldsymbol{\Omega}_x(\boldsymbol{I} + \boldsymbol{\Sigma}_u\boldsymbol{\Omega}_x)^{-1}\boldsymbol{\Sigma}_u\boldsymbol{\Omega}_x$ is equal to a zero matrix. This is generally not the case when the matrix $\boldsymbol{\Sigma}_u$ is not zero.

When no measurement error is present, i.e. the $\boldsymbol{x}_i$ are directly observed, the sample covariance matrix $\boldsymbol{S} = n^{-1}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$, with $\bar{\boldsymbol{x}}$ being the sample mean, is a consistent estimator for $\boldsymbol{\Sigma}_x$. However it has the rank of at

most $n < d$, so it is not invertible to estimate Ω_x. When measurement error is present, we assume the covariance matrix of measurement error Σ_u is known or estimable from replicates. A naïve approach is first to estimate Σ_x by $\tilde{\Sigma}_x = \boldsymbol{S}_w - \Sigma_u$, where $\boldsymbol{S}_w$ denotes the sample covariance from contaminated data $\boldsymbol{W}$, and then to invert $\tilde{\Sigma}_x$ to estimate Ω_x. The main issue with this approach is that $\tilde{\Sigma}_x$ is generally not positive definite. This implies its inverse is also not positive definite, which is necessary to find a consistent estimate Ω_x. Hence, a correction procedure to estimate $\boldsymbol{\Omega_x}$ need not rely upon the sample covariance matrix $\tilde{\Sigma}_x$ directly. Furthermore, the procedure should also be able to incorporate sparsity constraints to recover the graphical model structure. These requirements are addressed by the procedure described in the next section.

3. The IRO-BAGUS algorithm

In a recent work, Liang et al. (2018) develop a methodology to efficiently handle high dimensional problems with missing data. Their solution is an EM-algorithm variant which alternates between two steps, the imputation step and regularized optimization step; we refer to their algorithm as the IRO algorithm. Denote the missing data as Y, and observed data as X. Also denote the desired parameter to be estimated by θ, and begin with some initial guess $\theta^{(0)}$. During the tth iteration, the IRO algorithm generates Y from the distribution given by the current estimate of θ, i.e. $Y \sim \pi(Y|X, \theta^{(t-1)})$. Then, using X and Y, maximizes θ, under regulation, using the full likelihood. Liang et al. (2018) show that this procedure results in a consistent estimate of $\theta^{(t)}$, and results in a Markov chain with stationary distribution.

We make use of this framework for our current problem pertaining to mismeasured observations instead of missing values. The problems are naturally related in the sense that both are generating values of the true process from some estimated underlying distribution. We return to the hierarchical structure of the problem, i.e. $\boldsymbol{w}|\boldsymbol{x}, \Sigma_u \sim N_d(\boldsymbol{x}, \boldsymbol{\Sigma_u})$ and $\boldsymbol{x} \sim N_d(\boldsymbol{0}, \boldsymbol{\Sigma_x})$. The IRO algorithm proceeds iteratively between the two following steps:

- *Imputation step*: At iteration t, draw $\boldsymbol{X}^{(t)} = (\boldsymbol{x}_1^{(t)}, \ldots, \boldsymbol{x}_d^{(t)})$ from the posterior full conditional of $\boldsymbol{X}$, using the current estimate of $\boldsymbol{\Omega_x}^{(t-1)}$. Specifically, for $i = 1, \ldots, n$, draw $\boldsymbol{x}_i^{(t)}|\boldsymbol{w}, \boldsymbol{\Omega_x}^{(\boldsymbol{t-1})} \sim N_d(\boldsymbol{\Lambda^{-1}\Omega_u w}_i, \boldsymbol{\Lambda}^{-1})$ where $\boldsymbol{\Lambda} = (\boldsymbol{\Omega_x}^{(t-1)} + \boldsymbol{\Omega_u})$. Note that the posterior distribution of $\boldsymbol{x}_i$ depends only on $\boldsymbol{w}_i$ due to independence. This allows for easy generation of data from the true underlying distribution.
- *Regularization Step*: Apply a regularization to the generated $\boldsymbol{X}^{(t)}$ and obtain a new estimate of $\boldsymbol{\Omega_x}^{(t)}$.

In this work, the regularization step is carried out based on a recent Bayesian methodology, called BAGUS. Hence, the whole algorithm is referred to as the IRO-BAGUS algorithm. We briefly note that the algorithm proposed relies on the, typically, unknown measurement error covariance Σ_u. Assume replicates are observed and an independence assumption is made between the variables, then the diagonal of Σ_u is estimable, as elaborated in Appendix. The next Sections 3.1–3.3 outline prior specifications, the full model, and variable selection for BAGUS. After that, Section 3.4 discusses consistency of the IRO-BAGUS estimate.

3.1. The spike-and-slab lasso prior specification

Let ω_{ij} denote the (i, j) element of Ω_x. Recently, a non-convex, continuous relaxation penalty for the spike-and-slab prior was created for the standard lasso problem (Ročková and George, 2018). This prior was extended to the case of graphical models by Gan et al. (2018), and is given by

$$\pi(\omega_{ij}) = \frac{\eta}{2v_1} \exp\left\{-\frac{|\omega_{ij}|}{v_1}\right\} + \frac{1-\eta}{2v_0} \exp\left\{-\frac{|\omega_{ij}|}{v_0}\right\} \tag{2}$$

for the off diagonal elements ($i \neq j$), where $0 < v_0 < v_1$ and $0 < \eta < 1$. This prior can be interpreted as a mixture of the spike-and-slab prior. The first component of the mixture has prior probability η, and is associated with the slab component, i.e. $\omega_{ij} \neq 0$. Conversely, with prior probability $1 - \eta$ the element is from the spike component, suggesting $\omega_{ij} = 0$.

Traditionally, the spike-and-slab prior has a point mass component at 0 and some other continuous distribution for the slab component. This is to represent setting unwanted terms exactly to 0. Here, both the spike and the slab components are distributed according to a Laplace distribution; both are centered at 0, but the spike is more tightly centered by a smaller variance term than the slab. This relaxation of the spike-and-slab prior allows for efficient gradient based algorithms, while still being theoretically sound as shown in Ročková et al. (2018).

Shrinkage is not desired on the diagonal elements, so a different weakly informative exponential prior is given instead, $\pi(\omega_{ii}) = \tau \exp\{-\tau\omega_{ii}\}$. Another consideration for the prior of Ω_x is to ensure the whole matrix to be positive definite, denoting as $\Omega_x \succ 0$. Moreover, in line with Gan et al. (2018), we require the spectral norm to be bounded above by some value B, $\|\Omega_x\| \leq B$. This assumption will be important for establishing consistency going forward. The full prior distribution for Ω_x is then given by

$$\pi(\Omega_x) = \prod_{i<j} \pi(\omega_{ij}) \prod_i \pi(\omega_{ii}) I(\Omega_x \succ 0) I(\|\Omega_x\| \leq B). \tag{3}$$

3.2. The full model

Without measurement error, the posterior distribution is specified as

$$\pi(\boldsymbol{\Omega}_x|\boldsymbol{X}) \propto \prod_{i=1}^{n} \pi(\boldsymbol{x}_i|\boldsymbol{\Omega}_x)\pi(\boldsymbol{\Omega}_x). \tag{4}$$

The full conditionals can be derived for (4), but, to avoid computationally expensive MCMC sampling for this large dimensional problem, Gan et al. (2018) opted to instead find the mode of the posterior distribution, often referred to as the MAP. The MAP can be found by minimizing the uncontaminated (UC) objective

$$L^{\text{UC}}(\boldsymbol{\Omega}_x) = \log \pi(\boldsymbol{\Omega}_x|\boldsymbol{X}) = \frac{n}{2}\left(\text{tr}(\boldsymbol{X}^T\boldsymbol{\Omega}_x\boldsymbol{X}) - \text{logdet}(\boldsymbol{\Omega}_x)\right) + \sum_{i<j}\pi(\omega_{ij}) + \sum_{i}\pi(\omega_{ii}) + K \tag{5}$$

with respect to $\boldsymbol{\Omega}_x$, where K is the normalizing constant in (4). To this end, Gan et al. (2018) proved the local convexity of (4) when $\|\boldsymbol{\Omega}_x\| \leq B < \infty$, which allows an easy optimization procedure that converges asymptotically to the correct precision matrix.

3.3. Variable selection

Many practitioners use Gaussian graphical models for the purpose of identifying non-zero entries of $\boldsymbol{\Omega}_x$, which signify conditional dependencies among the two different variables. The spike-and-slab lasso formulation allows for this quite easily by viewing the optimization as an instance of the EM-algorithm and defining the hierarchical prior

$$\begin{cases} \omega_{ij}|r_{ij} = 0 \sim \text{Laplace}(0, v_0) \\ \omega_{ij}|r_{ij} = 1 \sim \text{Laplace}(0, v_1). \end{cases} \tag{6}$$

Here, r_{ij} is the random indicator that the element of the precision matrix follows from the spike or the slab component, where $r_{ij} \sim \text{Bern}(\eta)$. A further hierarchical level can be added by treating η as random instead of a fixed hyperparameter. Recent work from Deshpande et al. (2017) illustrates this and is line with the spike-and-slab Lasso setting of Ročková and George (2018). Given our purpose is to study the effect of the measurement error, we choose to treat η as fixed.

The conditional posterior distribution for r_{ij} is also Bernoulli, with probability of success

$$p_{ij} = \frac{v_1}{v_0}\frac{1-\eta}{\eta}\exp\left\{|\omega_{ij}|\left(\frac{1}{v_1} - \frac{1}{v_0}\right)\right\}. \tag{7}$$

We will use the MAP estimate of ω_{ij} in (7) to calculate p_{ij} and use it as the approximate probability of inclusion. A hard threshold T will need to be specified for the inclusion probability matrix to select the final model. Let $\boldsymbol{R}$ and $\boldsymbol{P}$ denote the matrix whose elements are r_{ij} and p_{ij} respectively for $i = 1, \ldots, p,\ j = 1, \ldots, p$; then, to obtain the final model, the (i, j) element of $\boldsymbol{\Omega}_x$ is set to zero if the MAP of p_{ij} is less than T. In practice, it might be better to forgo this inclusion threshold, and instead rank-order the p_{ij} for purposes of downstream investigation. This can enable practitioners to consider more practical aspects of graphical model and application at hand, such as network interpretability and stability,

3.4. Consistency of the IRO-BAGUS algorithm

The entire data generation process for the contaminated sample is summarized below:

$$\begin{aligned}
&\boldsymbol{w}_i|\boldsymbol{x}_i, \boldsymbol{\Omega}_x \sim N_d(\boldsymbol{x}_i, \boldsymbol{\Sigma}_u),\\
&\boldsymbol{x}_i|\boldsymbol{\Omega}_x \sim N_d(\boldsymbol{0}, \boldsymbol{\Omega}_x^{-1}),\\
&\omega_{ij}|r_{ij} = 0, v_0 \sim \text{Laplace}(0, v_0),\\
&\omega_{ij}|r_{ij} = 1, v_1 \sim \text{Laplace}(0, v_1),\\
&\omega_{ii} \sim \text{Exp}(\tau),\\
&r_{ij}|\eta \sim \text{Bern}(\eta);
\end{aligned}$$

here, $i \neq j$ and $i, j = 1, \ldots, n$. Instead of approximating the posterior distribution of all the parameters, the IRO-BAGUS algorithm iteratively generates realizations of uncontaminated data, $\boldsymbol{X}$, then optimizes $\boldsymbol{\Omega}_x$ with these generated values. Under some technical conditions, the IRO algorithm is shown to produce a consistent estimate after each iteration in the context of missing data when the regularization step results in a consistent estimate (Liang et al., 2018). We show that these conditions are also held in the case of contaminated data, so the IRO-BAGUS algorithm results in a consistent estimate. Theorem 1 is the analogue statement of consistency as in the missing data case. The proof is given in the appendix.

Theorem 1. *With the model assumptions as shown above and assuming* $\|\boldsymbol{\Omega}_x\| \leq B$, *then the estimate* $\boldsymbol{\Omega}_x^{(t)}$ *is uniformly consistent to* $\boldsymbol{\Omega}_x$ *when* $\log(t) = \mathcal{O}(n)$.

It can be seen that the nature of the IRO algorithm is similar to that of MCMC. With additional mild conditions, Liang et al. (2018) note that the IRO results in a Markov chain with a stationary distribution, and hence the average of the maximization steps are consistent estimates of the underlying parameters. Our final estimates are the averaged regularized optimization steps given by BAGUS from the imputed data at each iteration, removing a small number of the beginning iterations as burn-in. By averaging instead of taking only the final iteration, we make the analysis less variable. In this sense, the relationships that the correction procedure identifies are more likely to be true relationships, cutting down on the number of false positives.

3.5. Parameter tuning

Four hyperparameters are involved in BAGUS, η, τ, v_0, and v_1, and there is one threshold T that needs to be specified to determine the final model. For the hyperparameters, as with Gan et al. (2018), we always set $\eta = 0.5$ and $\tau = v_0$, which leaves two hyperparameters to tune. For the threshold T, we consider 30 equi-distant values between 0 and 1. To choose a final model, we follow Gan et al. (2018), who suggest a BIC-like criteria to select the best model from a grid of hyperparameters. This criteria is

$$\text{BIC} = n(\text{tr}(\boldsymbol{S}\hat{\boldsymbol{\Omega}}_x) - \text{logdet}(\hat{\boldsymbol{\Omega}}_x)) + \log(n) \times q,$$

where $\hat{\boldsymbol{\Omega}}_x$ is the estimated precision matrix and q is the number of non-zero elements of the estimated in the upper diagonal of the precision matrix. We use this in similar fashion for the IRO procedure, but instead we use the averaged $\boldsymbol{\Omega}_x^{(t)}$ in the BIC calculation.

4. Simulation study

4.1. Simulation setup

We investigate the performance of our methodology under several different settings. For each setting we generate $\boldsymbol{x}_i$ following a d-variate Gaussian distribution with mean $\boldsymbol{0}$ and precision matrix $\boldsymbol{\Omega}_x$ according to some graphical structure; we refer to this as the *true data*. Then, the contaminated observations $\boldsymbol{w}_i$ were generated from $\boldsymbol{w}_i = \boldsymbol{x}_i + \boldsymbol{u}_i$, where $\boldsymbol{u}_i \sim N_d(\boldsymbol{0}, \boldsymbol{\Sigma}_u)$, $i = 1, \ldots, n$. The measurement error covariance matrix $\boldsymbol{\Sigma}_u$ is assumed to be a diagonal matrix, with element $[\boldsymbol{\Sigma}_u]_{ii} = \gamma\,[\boldsymbol{\Sigma}_x]_{ii}$, where $[\boldsymbol{\Sigma}_x]_{ii}$ is the variance of the dimension x^i. In other words, the constant γ controls the noise-to-signal ratio on each variable. For the purposes of simulation, we assume the amount of measurement error to be known.

To generate the true data we use the *huge* package (Zhao et al., 2012). We inspect two different types of graphs, referred to as *hub*, and *random*; we expand on these below where ω_{ij} denotes the (i, j) element of $\boldsymbol{\Omega}_x$.

1. Hub: The d variables are divided into $K = d/20$ non-overlapping groups, each group having 20 elements. The kth group has a "center" at $\boldsymbol{X}_{20(k-1)+1}$, which has connection to every other element in that group, $k = 1, \ldots, K$. In other words, $\omega_{ij} = \omega_{ji} = 1$ if and only if $i = 20(k-1)+1$, $j = i + m$, $1 \le m \le 20$, $k = 1, \ldots, K$. With that structure, the graph only has $d - K$ edges out of $d(d-1)/2$ possible edges.
2. Random: For $1 \le i < j \le d$, $\omega_{ij} = 1$ with probability $\frac{3}{d}$, 0 otherwise.

We illustrate the structures in Fig. 1.

Each model was generated with $n = 100$ observations. We inspect each model for $d = \{100, 200\}$ and $\gamma = \{0.1, 0.25, 0.5\}$. The amount of correction-imputations was set to be 50, with the first 20% discarded as burn-in; we note that we inspected 25 and 100 imputations with the same percentage of burn-in samples with minimal differences in output. Each setting was replicated 50 times, and the final results are the average of these replicates. We assume the measurement error to be known for the correction procedure. Hyperparameter tuning was done as described in Section 3.5. Because measurement error is often ignored in the context of GGMs, our simulations also provide perspective onto the negative effect that measurement error can impose on model performance.

To inspect model performance, we examine both the estimated precision matrix and the ability to do variable selection using BAGUS on the true data (true), BAGUS on the contaminated data (naïve), and our IRO-BAGUS methodology on the contaminated data (corrected). For each estimated precision matrix $\hat{\boldsymbol{\Omega}}_x$, estimation error is measured by $\|\hat{\boldsymbol{\Omega}}_x - \boldsymbol{\Omega}_x\|_F$, and variable selection, evaluated by different metrics involving the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), are reported: specificity (SPE), sensitivity (SEN), precision (PRE), accuracy (ACC), and Matthews correlation coefficient (MCC); these values are defined as

$$\text{SPE} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \qquad \text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$\text{PRE} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}},$$
$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}.$$

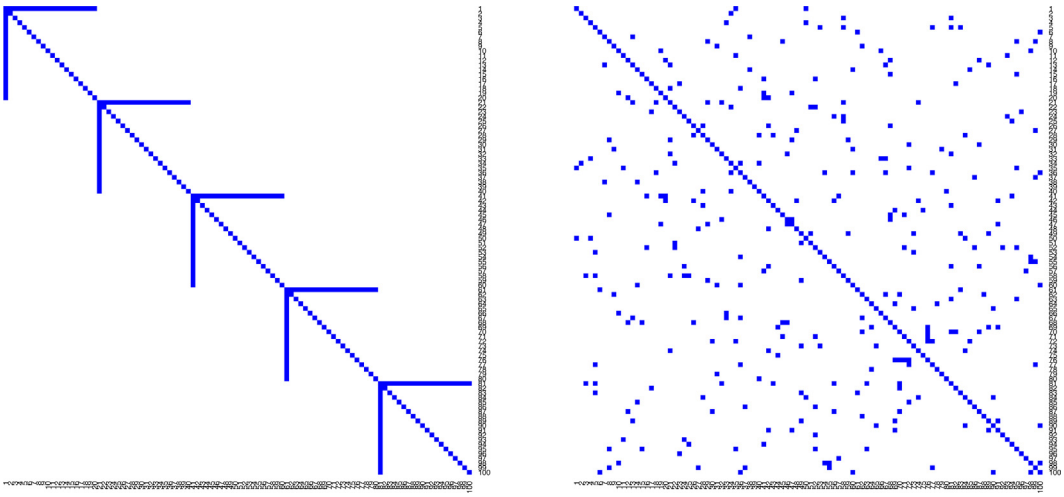

Fig. 1. Graphical representation for $d = 100$ of the hub (left) and random (right) structure, respectively. While the hub structure is fixed for a given d, the random graph is subject to change due to the generation process.

Table 1
Simulation results for the hub graph structure, as specified in Section 4.1. For each signal-to-noise ratio and d, the true, naïve, and corrected models are shown for metrics defined in Section 4.1.

γ	d	Model	SPE	SEN	PRE	ACC	MCC	FROB	AUC
0.1	100	True	0.99	0.75	0.72	0.99	0.72	4.92	0.83
		Naïve	0.98	0.71	0.45	0.98	0.55	6.58	0.80
		Corrected	0.99	0.63	0.66	0.99	0.64	5.80	0.82
	200	True	1.00	0.74	0.64	0.99	0.69	7.18	0.83
		Naïve	0.99	0.67	0.43	0.99	0.53	9.40	0.80
		Corrected	1.00	0.64	0.57	0.99	0.60	8.35	0.83
0.25	100	True	0.99	0.76	0.72	0.99	0.74	4.89	0.83
		Naïve	0.98	0.59	0.34	0.97	0.43	8.15	0.76
		Corrected	0.99	0.50	0.56	0.98	0.52	6.86	0.80
	200	True	1.00	0.74	0.64	0.99	0.68	7.31	0.83
		Naïve	0.98	0.59	0.25	0.98	0.38	11.57	0.77
		Corrected	0.99	0.49	0.48	0.99	0.48	9.90	0.81
0.5	100	True	0.99	0.76	0.70	0.99	0.72	4.98	0.83
		Naïve	0.98	0.40	0.26	0.97	0.31	9.52	0.71
		Corrected	0.99	0.28	0.54	0.98	0.37	7.94	0.76
	200	True	1.00	0.74	0.64	0.99	0.68	7.30	0.83
		Naïve	0.98	0.41	0.17	0.98	0.26	13.43	0.72
		Corrected	1.00	0.24	0.47	0.99	0.33	11.62	0.78

Additionally, we also report the area under the ROC curve (AUC). Since the AUC marginalizes over all possible thresholds and tuning parameters, it gives us insight into the amount of separation of the non-zero and zero elements for each method. This reflects the overall ability to recover the graph structure. These different metrics give insight into the tradeoffs and gains of each setting.

4.2. Simulation results

Tables 1 and 2 present the results for the hub and random structure, respectively.

To begin, we note the effect of the increasing measurement error. This can be observed by examining the growing difference in the performance of the true and naïve model when holding d fixed and increasing the amount of contamination. Focusing on the hub structure, a decrease in the quality of selection and estimation can be observed for each setting, which grows worse with more contamination; for example, when $d = 100$ the AUC drops from 0.83 to 0.80 for $\gamma = 0.1$, but drops from 0.83 to 0.76 for $\gamma = 0.5$. The estimated precision matrix from the naïve model grows worse with measurement error, and is about 50% worse when the signal-to-noise is $\gamma = 0.5$. Additionally, we note that the random structure is harder to recover than the hub structure; even without measurement error, the true estimator performs worse regarding both selection and estimation to recover the random structure.

Table 2
Simulation results for the random graph structure, as specified in Section 4.1. For each signal-to-noise ratio and d, the true, naïve, and corrected models are shown for metrics defined in Section 4.1.

γ	d	Model	SPE	SEN	PRE	ACC	MCC	FROB	AUC
0.1	100	True	0.98	0.67	0.58	0.98	0.61	5.19	0.81
		Naïve	0.97	0.64	0.40	0.96	0.48	5.59	0.79
		Corrected	0.99	0.46	0.68	0.98	0.54	5.25	0.81
	200	True	0.99	0.56	0.48	0.98	0.50	8.13	0.79
		Naïve	0.99	0.40	0.53	0.99	0.45	7.57	0.77
		Corrected	1.00	0.36	0.60	0.99	0.45	7.26	0.80
0.25	100	True	0.99	0.65	0.58	0.98	0.60	5.21	0.81
		Naïve	0.98	0.47	0.41	0.96	0.42	6.35	0.76
		Corrected	0.99	0.29	0.64	0.97	0.42	5.68	0.78
	200	True	0.99	0.58	0.45	0.98	0.50	8.31	0.80
		Naïve	0.99	0.38	0.34	0.98	0.34	8.78	0.74
		Corrected	1.00	0.26	0.52	0.99	0.36	8.02	0.78
0.5	100	True	0.99	0.65	0.58	0.98	0.60	5.20	0.81
		Naïve	0.98	0.33	0.34	0.96	0.31	7.61	0.71
		Corrected	1.00	0.13	0.59	0.97	0.27	6.55	0.73
	200	True	0.99	0.58	0.46	0.98	0.50	8.38	0.80
		Naïve	0.98	0.29	0.20	0.97	0.23	10.34	0.69
		Corrected	1.00	0.10	0.46	0.98	0.21	9.25	0.72

Next, we compared the performance of the best model obtained for each method using the tuning method as described in Section 3.5. Regarding the quality of the estimated precision matrix, measured by the Frobenius norm of the difference of the estimated and true precision matrix, in both settings the corrected model outperforms the naïve model's estimate of the precision matrix. In the hub structure this improvement is often of the order of 15%, while in the random structure a 10% improvement is generally observed.

Regarding selection performance, compared to the corrected estimator, the naïve estimator has a higher sensitivity and lower specificity, meaning that it does find more true relationships at the expense of identifying more false positives. Note that because the true graph is sparse, a 1% decrease in specificity already results in many false positives (for instance, with $d = 100$ and the hub structure, the number of false negatives is 4855, so a 1% decrease in specificity results in about 50 more false positives). As a result, when considering selection metrics that take both true positives and false positives into account (such as precision, accuracy, MCC), the corrected still outperforms the naïve estimator in almost all the considered settings. For example, regarding precision (PRE), which measures the correctness of the identified positives by the model, the corrected model outperforms the naïve model in every setting by at least a factor of 1.5. When the measurement error variance grows, the precision of the corrected estimator can sometimes be more than 2 times better than the naïve estimator. In addition, the MCC shows preferable performance to the corrected model in the hub graph structure; comparable MCC results are observed in the random graph structure.

Finally, we turn to the overall performance of each method in recovering the graphical model, as measured by the AUC. It is observed the corrected model outperforms the naïve model in every setting. This suggests an overall improved model from the IRO correction. Intuitively, incorporating information across each of the IRO iterations resulted in an analysis that favored identifying relationships that were more certain. In practice, practitioners can also use complementary information from the estimated precision matrix for downstream analysis to choose the final model. This enables practitioners to take other practical considerations into account, such as complexity or connectivity of the graph.

5. Data analysis

A common source of noise in analysis involving gene expression datasets is measurement error (Rocke and Durbin, 2001). Gaussian graphical models have been employed to inspect the relationship of different genes in varying experiments (Krämer et al., 2009). We illustrate our methodology using an Affymetrix microarray dataset containing 144 subjects of favorable histology Wilms tumors hybridized to the Affymetrix Human Genome U133A Array (Huang et al., 2009). The data is publicly available on the GEO website, dataset GSE10320 uploaded 1/30/2009. A feature of Affymetrix data, and many other gene expression measurement platforms, is the use of multiple probes for each gene for each patient, giving replicate measurements for each patient's gene measurement. The replicates for each patient enable an estimate of the measurement error, where we again assume the amount of contamination is independent across genes.

We follow the preprocessing steps taken in Sørensen et al. (2015) and Nghiem and Potgieter (2018), which applied their methods to this study in the context of measurement error in variable selection for linear models. The process begins by processing the raw data with the Bayesian Gene Expression (BGX) package (Turro et al., 2007). BGX creates a posterior distribution for the log-scale expression level of each gene in each sample. The study recorded measurements for 22 283 different genes.

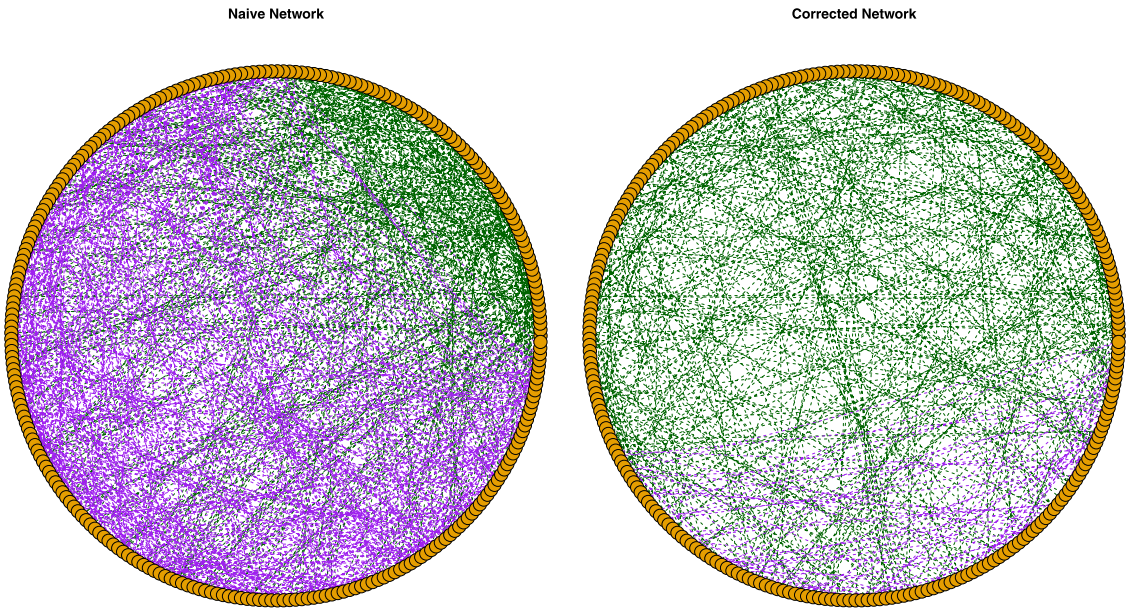

Fig. 2. The conditional pair-wise relationships for each of the 273 genes remaining after filtering from the Wilms tumor study. Each edge represents a conditional pair-wise dependency between two nodes. The left shows the naïve analysis, not correcting for measurement error, and the right shows the corrected analysis, correcting for measurement error. Green edges signify edges found on both graphs, and purple signifies analysis specific edges. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To remove unnecessary computational burden, we reduced the number of genes by applying four different filters in the following order. The first filter removes expression values that do not have a corresponding Entrez gene ID in the NCBI database (O'Leary et al., 2015). The second filter removes minimally expressed genes by requiring at least 25% of samples to have intensities above 100 fluorescence units. The third filter removes expression values with low variability by requiring the interquartile range to be at least 0.6 on the log scale. The last filter removes expression values that have an error to signal-to-noise ratio greater than 0.5, which we discuss in more depth below. After filtering, there were 273 expression values remaining for the analysis.

Now, we discuss how we estimate the measurement error of each gene. We assume that the measurement error variance is constant across patients for a given gene. We also assume that the measurement error is independent for each gene, and need not be equal for each gene. Let $\hat{\boldsymbol{\mu}} = (\hat{\mu}_{1j}, \ldots, \hat{\mu}_{nj})^T$ denote the estimated vector of the patient's gene expression levels for gene j. Further, let $\bar{\mu} = n^{-1}\sum_{j=1}^{n} \hat{\mu}_{ij}$ and $\hat{\sigma}_j^2 = n^{-1}\sum_{j=1}(\hat{\mu}_{ij} - \bar{\mu}_j)^2$ denote the mean and variance of each gene, respectively. For patient i, standardized measurements are given by $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{ip})$, calculated as $W_{ij} = \hat{\sigma}_j^{-1}(\hat{\mu}_{ij} - \bar{\mu}_j)$ for each $j = 1, \ldots, 273$.

Let $\text{var}(\hat{\mu}_{ij})$ denote the posterior variance of the estimated distribution of patient i's gene j. These estimates are then combined as $\hat{\sigma}_{u,j}^2 = n^{-1}\sum_{i=1}^{n} \text{var}(\hat{\mu}_{ij})$. The measurement error covariance matrix of the standardized data $\boldsymbol{W}$ is then estimated by diagonal matrix $\hat{\boldsymbol{\Sigma}}_u$, where $(\hat{\boldsymbol{\Sigma}}_u)_{j,j} = \hat{\sigma}_{u,j}^2/\hat{\sigma}_j^2$ for $j = 1, \ldots, p$ and off-diagonal elements are 0. The fourth filter can be now formalized, where genes are removed if $\hat{\sigma}_{u,j}^2 \geq 0.5\hat{\sigma}_j^2$; i.e. only genes with a noise-to-signal ratio less than 1 are kept for the analysis.

The original BAGUS algorithm and the IRO-BAGUS algorithm were run for the genes remaining after filtering. We illustrate the conditional pair-wise dependencies of the genes in Fig. 2. The naïve analysis is shown on the left and the corrected on the right, where the green edges signify relationships found by both procedures and purple edges signify procedure specific relationships. As with the simulations, the corrected BAGUS found fewer conditional pair-wise relationships; for this data set, the original BAGUS and IRO-BAGUS found 1045 and 552 conditional pair-wise relationships, respectively. Of the 1045 naïve pair-wise relationships, 42% were also found in the corrected pair-wise relationships; similarly, of the 552 corrected conditional pair-wise relationships, 80% were found in the naïve model. The large percentage overlap of relationships in the corrected model with relationships in the naïve model suggests that most relationships in the corrected model are true relationships. Conversely, the small percentage overlap of relationships in the naïve model with those in the correct model suggests that the naïve model is finding many false positive relationships.

In addition, we inspect the quality of the estimates of each precision matrix. The Frobenius norm of the naïve model was 59.30 for 1045 pairs, whereas the Frobenius norm for the corrected model was 46.83 for 552 pairs. This immediately suggests that the corrected estimate is more certain of the identified relationships existing, given the magnitude per element is much greater for the corrected method. The median magnitude of non-zero elements is 0.95 and 1.36 for the naïve and corrected methods, respectively. When filtering down to overlapping non-zero elements, we find the median

magnitude to be 1.22 and 1.42 for the naïve and corrected, respectively. Although the corrected method possibly misses some possible relationships, we get stronger signal for the relationships that are present. This occurrence has been also noted in the previous literature of measurement error correction in high dimensional settings, such as Sørensen et al. (2015) and Nghiem and Potgieter (2018).

6. Conclusion

We proposed a correction methodology for Gaussian graphical models when contaminated with additive measurement error. The core solution to the problem involves using the imputation–regularization algorithm to generate the true values of underlying process with a consistent estimate of the precision matrix. This provides a consistent, positive-definite estimate of the true precision matrix, which, as simulations illustrate, removes many false positive pair-wise relationships. Additionally, we show marked improvements in the AUC of the threshold matrix, indicating better separation of the underlying relationships. From a practitioner's point of view, this allows for more reliable downstream analysis and a stronger set of results from which to continue research.

To our knowledge, the novel imputation–regularization algorithm has yet to be used for problems pertaining to contaminated data. This provides an avenue of future research on measurement error in high-dimensional applications, which is starting to gain attention. Moreover, many practical issues still remain in the Gaussian graphical model context, such as the tuning of hyperparameters and the interpretation of the output from the Gibbs sampler-like IRO algorithm. Another potential avenue of research to pursue is when the amount of measurement error is unknown and not assumed independent. In this case, sparsity would need to be imposed on $\mathbf{\Omega}_u$ in conjunction with Ω_x, posing a challenging, but useful, computational procedure.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2020.107085.

References

Cai, Tony, Liu, Weidong, Luo, Xi, 2011. A constrained l1 minimization approach to sparse precision matrix estimation. J. Amer. Statist. Assoc. 106 (494), 594–607.

Carroll, Raymond J, Ruppert, David, Crainiceanu, Ciprian M, Stefanski, Leonard A, 2006. Measurement Error in Nonlinear Models: a Modern Perspective. Chapman and Hall/CRC.

Dempster, Arthur P., 1972. Covariance selection. Biometrics 157–175.

Deshpande, Sameer K., Rockova, Veronika, George, Edward I., 2017. Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. arXiv preprint arXiv:1708.08911.

Friedman, Jerome, Hastie, Trevor, Tibshirani, Robert, 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9 (3), 432–441.

Gan, Lingrui, Narisetty, Naveen N., Liang, Feng, 2018. Bayesian regularization for graphical models with unequal shrinkage. J. Amer. Statist. Assoc. 1–14.

Huang, Chiang-Ching, Gadd, Samantha, Breslow, Norman, Cutcliffe, Colleen, Sredni, Simone T, Helenowski, Irene B, Dome, Jeffrey S, Grundy, Paul E, Green, Daniel M, Fritsch, Michael K, et al., 2009. Predicting relapse in favorable histology wilms tumor using gene expression analysis: a report from the renal tumor committee of the children's oncology group. Clin. Cancer Res. 15 (5), 1770–1778.

Johnstone, Iain M., et al., 2001. On the distribution of the largest eigenvalue in principal components analysis. Ann. Statist. 29 (2), 295–327.

Khare, Kshitij, Oh, Sang-Yun, Rajaratnam, Bala, 2015. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. J. R. Stat. Soc. Ser. B Stat. Methodol. 77 (4), 803–825.

Krämer, Nicole, Schäfer, Juliane, Boulesteix, Anne-Laure, 2009. Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC Bioinformatics 10 (1), 384.

Lauritzen, Steffen L., 1996. Graphical Models, Vol. 17. Clarendon Press.

Liang, Faming, Jia, Bochao, Xue, Jingnan, Li, Qizhai, Luo, Ye, 2018. An imputation–regularized optimization algorithm for high dimensional missing data problems and beyond. J. R. Stat. Soc. Ser. B Stat. Methodol. 80 (5), 899–926.

Liu, Han, Wang, Lie, et al., 2017. Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. Electron. J. Stat. 11 (1), 241–294.

Nghiem, Linh, Potgieter, Cornelis, 2018. Simulation-selection-extrapolation: Estimation in high-dimensional errors-in-variables models. arXiv preprint arXiv:1808.10477.

O'Leary, Nuala A, Wright, Mathew W, Brister, J Rodney, Ciufo, Stacy, Haddad, Diana, McVeigh, Rich, Rajput, Bhanu, Robbertse, Barbara, Smith-White, Brian, Ako-Adjei, Danso, et al., 2015. Reference sequence (refseq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44 (D1), D733–D745.

Petersen, Kaare Brandt, Pedersen, Michael Syskind, et al., 2008. The matrix cookbook. Tech. Univ. Denmark 7 (15), 510.

Rocke, David M., Durbin, Blythe, 2001. A model for measurement error for gene expression arrays. J. Comput. Biol. 8 (6), 557–569.

Ročková, Veronika, George, Edward I., 2018. The spike-and-slab lasso. J. Amer. Statist. Assoc. 113 (521), 431–444.

Ročková, Veronika, et al., 2018. Bayesian estimation of sparse signals with a continuous spike-and-slab prior. Ann. Statist. 46 (1), 401–437.

Segal, Eran, Shapira, Michael, Regev, Aviv, Pe'er, Dana, Botstein, David, Koller, Daphne, Friedman, Nir, 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. 34 (2), 166.

Sørensen, Øystein, Frigessi, Arnoldo, Thoresen, Magne, 2015. Measurement error in lasso: Impact and likelihood bias correction. Statist. Sinica 809–829.

Tan, Kean Ming, Ning, Yang, Witten, Daniela M., Liu, Han, 2016. Replicates in high dimensions, with applications to latent variable graphical models. Biometrika 103 (4), 761–777.

Turro, Ernest, Bochkina, Natalia, Hein, Anne-Mette K, Richardson, Sylvia, 2007. BGX: a Bioconductor package for the bayesian integrated analysis of affymetrix genechips. BMC Bioinformatics 8 (1), 439.
Yuan, Ming, Lin, Yi, 2007. Model selection and estimation in the Gaussian graphical model. Biometrika 94 (1), 19–35.
Zakharkin, Stanislav O, Kim, Kyoungmi, Mehta, Tapan, Chen, Lang, Barnes, Stephen, Scheirer, Katherine E, Parrish, Rudolph S, Allison, David B, Page, Grier P, 2005. Sources of variation in affymetrix microarray experiments. BMC Bioinform. 6 (1), 214.
Zhao, Tuo, Liu, Han, Roeder, Kathryn, Lafferty, John, Wasserman, Larry, 2012. The huge package for high-dimensional undirected graph estimation in R. J. Mach. Learn. Res. 13 (Apr), 1059–1062.

Further reading

Honorio, Jean, 2012. Lipschitz parametrization of probabilistic graphical models. arXiv preprint arXiv:1202.3733.