# QUASI-BAYESIAN ESTIMATION OF LARGE GAUSSIAN GRAPHICAL MODELS

YVES F. ATCHADÉ

(Dec. 2018; first draft Dec. 2015)

ABSTRACT. This paper deals with the Bayesian estimation of large precision matrices in Gaussian graphical models. We develop a quasi-Bayesian implementation of the neighborhood selection method of Meinshausen and Buhlmann (2006). The method produces a product-form quasi-posterior distribution that can be efficiently explored by parallel computing. Under some restrictions on the true precision matrix, we show that the quasi-posterior distribution contracts in the spectral norm at the rate of $O\left(s_\star\sqrt{\frac{\log(p)}{n}}\right)$, where $p$ is the number of nodes in the graph, $n$ the sample size, and $s_\star$ is the maximum degree of the undirected graph defined by the true precision matrix. We develop a Markov Chain Monte Carlo algorithm for approximate computations, following an approach from Atchadé (2015). We illustrate the methodology using real and simulated data examples.

## 1. INTRODUCTION

We consider the problem of fitting large Gaussian graphical models from limited data. More precisely, our goal is to estimate a sparse precision matrix $\vartheta \in \mathcal{M}_p^+$ from $p$-dimensional Gaussian observations $y^{(i)} \in \mathbb{R}^p$, $i = 1, \ldots, n$, where $\mathcal{M}_p^+$ denotes the cone of $\mathbb{R}^{p\times p}$ of symmetric positive definite matrices. The frequentist approach to this problem has generated an impressive literature over the last decade or so (see for instance Bühlmann and van de Geer (2011); Hastie et al. (2015) and the reference therein).

There is an interest, particularly in biomedical research, for statistical methodologies that can allow practitioners to incorporate external information in fitting such graphical models (Mukherjee and Speed (2008); Peterson et al. (2015)). This problem naturally calls for a Bayesian formulation and significant progress has been made in recent years (Dobra et al. (2011); Lenkoski and Dobra (2011); Khondker et al. (2013);

2000 *Mathematics Subject Classification.* 60F15, 60G42.

*Key words and phrases.* Gaussian graphical models, Quasi-Bayesian inference,pseudo-likelihood, posterior contraction, forward-backward approximation, Markov Chain Monte Carlo.

Y. F. Atchadé: University of Michigan, 1085 South University, Ann Arbor, 48109, MI, United States. *E-mail address:* yvesa@umich.edu.

Peterson et al. (2015); Banerjee and Ghosal (2015)). Another appealing aspect of the Bayesian framework is that it synthesizes all the available information on the parameter into a probability distribution for easy uncertainty quantification. However, most existing Bayesian methods for fitting graphical models do not scale well with the number of nodes in the graph. The main difficulty is computational, and hinges on the ability to handle interesting prior distributions on $\mathcal{M}_p^+$ when $p$ is large. The most commonly used class of priors distributions for Gaussian graphical models is the class of G-Wishart distributions (Atay-Kayis and Massam (2005)). However G-Wishart distributions have intractable normalizing constants, and become impractical for inferring large graphical models, due to the cost of approximating the normalizing constants (Dobra et al. (2011); Lenkoski and Dobra (2011)). Following the development of the Bayesian lasso of Park and Casella (2008) and other Bayesian shrinkage priors for linear regressions (Carvalho et al. (2010)), several authors have proposed prior distributions on $\mathcal{M}_p^+$ obtained by putting conditionally independent shrinkage priors on the entries of the matrix, subject to a positive definiteness constraint (Khondker et al. (2013)). However this approach does not give a direct estimation of the graph structure, which in many applications is the key quantity of interest. Furthermore, dealing with the positive definiteness constraint in the posterior distribution requires careful MCMC design, and becomes a limiting factor for large $p$.

The above discussion suggests that when dealing with large graphical models, some form of approximation is inescapable. Building on Atchade (2017), we propose a quasi-Bayesian approach for fitting large Gaussian graphical models using the pseudo-likelihood function that underpins the neighborhood selection method of Meinshausen and Buhlmann (2006). This choice gives a quasi-posterior distribution $\check{\Pi}_{n,p}$ that factorizes, and leads to a drastic improvement in the computing time needed for MCMC computation when a parallel computing architecture is used. We illustrate the method in Section 4 using simulated data where the number of nodes in the graph is $p \in \{100, 500, 1000\}$.

The idea of replacing the likelihood function by a pseudo-likelihood function is well-known. We refer the reader to Varin et al. (2011) for an in-depth discussion in the fixed-dimensional setting. The basic idea behind the use of pseudo-likelihood functions is to approximate a statistical model by a set of small-dimensional sub-models – typically constructed from conditional distributions. Each sub-model identifies only a small piece of the parameter of interest. A product of these sub-models is then used to identify the full-parameter. The idea is similar to inference by moment conditions (Li and Jiang (2014)), and is known to be a robust modeling approach, since only the sub-models are specified. In fixed-dimensional classical statistics, inference using

pseudo-models is known to be consistent, with the same $\sqrt{n}$ convergence rate as the maximum full-likelihood estimator (see Varin et al. (2011)). In this context, the price to pay for using a pseudo-model estimator is typically a larger asymptotic variance. In that sense, quasi-likelihood inference gives an approach to strike a better trade-off between robustness and computational tractability on one side, and statistical accuracy on the other.

We study the contraction properties of the quasi-posterior distribution $\check{\Pi}_{n,p}$ as $n, p \to \infty$. Under some restrictions on the true precision matrix, we show that $\check{\Pi}_{n,p}$ contracts in the spectral norm at the rate of $s_\star \sqrt{\frac{\log(p)}{n}}$ (see Theorem 7 and (16) for a precise statement), where $s_\star$ the maximum degree in the un-directed graph defined by the true precision matrix. The condition on the sample size $n$ for the results mentioned above to hold is $n \geq O\left(s_\star \log(p)\right)$, which shows that the quasi-posterior distribution can recover the true precision matrix, even in cases where $p$ exceeds $n$. The rate matches the frequentist rate of neighbood selection obtained in (Sun and Zhang (2013)). A full likelihood inference of $\vartheta$ yields the convergence rate of $\sqrt{\frac{S \log(p)}{n}}$, where $S$ is the number of non-zero entries of the true precision matrix. The full-likelihood rate was derived in the frequentist setting by Rothman et al. (2008), and in the Bayesian setting by Banerjee and Ghosal (2015). Note that typically, $S \sim p$. Hence, these rates seem to highlight an interesting high-dimensional phenomenon where a quasi-model converges at a faster rate than a full-likelihood inference, in addition to yielding computationally faster procedures.

The rest of the paper is organized as follows. Section 2 provides a general discussion of quasi-models and quasi-Bayesian inference. We specialized the discussion to Gaussian graphical models in Section 3. The theoretical analysis focuses on the Gaussian case, and is presented in Section 3, but the proofs are postponed to Section 5. The numerical experiments are presented in Section 4. A `MATLAB` implementation of the method is available from the author's website.

## 2. Quasi-Bayesian inference of graphical models

For integers $p \geq 1$, and $i \in \{1, \ldots, p\}$, let $\mathsf{Y}_i$ be a nonempty subset of $\mathbb{R}$, and set $\mathsf{Y} \stackrel{\text{def}}{=} \mathsf{Y}_1 \times \cdots \times \mathsf{Y}_p$, that we assume is equipped with a reference sigma-finite product measure $\mathrm{d}y$. We first consider a class of Markov random field distributions $\{f_\omega,\ \omega \in \Omega\}$ for joint modeling of $\mathsf{Y}$-valued random variables. Let $\mathcal{M}_p$ denote the set of all real symmetric $p \times p$ matrices equipped with the inner product $\langle A, B \rangle_{\mathsf{F}} \stackrel{\text{def}}{=} \sum_{i \leq j} A_i B_{ij}$, and norm $\|A\|_{\mathsf{F}} \stackrel{\text{def}}{=} \sqrt{\langle A, A \rangle_{\mathsf{F}}}$. As above, $\mathcal{M}_p^+$ denotes the subset of $\mathcal{M}_p$ of positive definite matrices. For $i = 1, \ldots, p$, and $1 \leq j < k \leq p$, let $B_i : \mathsf{Y}_i \to \mathbb{R}$ and $B_{jk} : \mathsf{Y}_j \times \mathsf{Y}_k \to \mathbb{R}$ be non-zero measurable functions that we assume known.

From these functions we define a $\mathcal{M}_p$-valued function $\bar{B}:\ \mathsf{Y}\to\mathcal{M}_p$ by

$$(\bar{B}(y))_{ij}=\begin{cases} B_i(y_i) & \text{if } i=j,\\ B_{ij}(y_i,y_j) & \text{if } i<j,\\ B_{ji}(y_j,y_i) & \text{if } j<i.\end{cases}$$

These functions define the parameter space

$$\Omega\stackrel{\text{def}}{=}\left\{\omega\in\mathcal{M}_p:\ Z(\omega)\stackrel{\text{def}}{=}\int_{\mathsf{Y}}e^{-\left\langle\omega,\bar{B}(y)\right\rangle_{\mathsf{F}}}\mathrm{d}y<\infty\right\}.$$

We assume that $\mathsf{Y}$ and $\bar{B}$ are such that $\Omega$ is non-empty, and we consider the exponential family $\{f_\omega,\ \omega\in\Omega\}$ of densities $f_\omega$ on $\mathsf{Y}$ given by

$$f_\omega(y)=\exp\left(-\left\langle\omega,\bar{B}(y)\right\rangle_{\mathsf{F}}-\log Z(\omega)\right),\ \ y\in\mathsf{Y}. \tag{1}$$

The model $\{f_\omega,\ \omega\in\Omega\}$ can be useful to capture the dependence structure between a set of $p$ random variables taking values in $\mathsf{Y}$. If $(Y_1,\ldots,Y_p)\sim f_\omega$, then the parameter $\omega$ encodes the conditional independence structure among the $p$ variables $(Y_1,\ldots,Y_p)$. In particular for $i\neq j$, $\omega_{ij}=0$ means that $Y_i$ and $Y_j$ are conditionally independent given all other variables. The random variables $(Y_1,\ldots,Y_p)$ can then be represented by an undirect graph where there is an edge between $i$ and $j$ if and only if $\omega_{ij}\neq 0$. This type of models are very useful in practice to tease out direct and indirect connections between sets of random variables. The version posited in (1) can accommodate mixed measurements where some of the $y_i$ take discrete values while other take continuous values.

**Example 1** (Gaussian graphical models)**.** One recovers the Gaussian graphical model by taking $\mathsf{Y}_i=\mathbb{R}$, $B_i(x)=x^2/2$, $B_{ij}(x,y)=xy$, $i<j$. In this case $\mathsf{Y}=\mathbb{R}^p$ equipped with the Lebesgue measure, and $\Omega=\mathcal{M}_p^+$.

**Example 2** (Potts models)**.** For integer $M\geq 2$, one recovers the $M$-states Potts model by taking $\mathsf{Y}_i=\{1,\ldots,M\}$. In this case, $\mathsf{Y}=\{1,\ldots,M\}^p$ equipped with the counting measure. Since $\mathsf{Y}$ is a finite set, we have $\Omega=\mathcal{M}_p$. An important special case of the Potts model is a version of the Ising model where $M=2$, and $B_i(x)=x$, and $B_{ij}(x,y)=xy$.

Suppose that we observe data $y^{(1)},\cdots,y^{(n)}$ where $y^{(i)}=(y_1^{(i)},\ldots,y_p^{(i)})'\in\mathsf{Y}$ is viewed as a column vector. We set $x\stackrel{\text{def}}{=}[y^{(1)},\ldots,y^{(n)}]'\in\mathbb{R}^{n\times p}$. Given a prior distribution $\Pi$ on $\Omega$, and given the data $x$, the resulting posterior distribution for learning $\omega$ is

$$\Pi_n(A|x)=\frac{\int_A\prod_{i=1}^n f_\omega(y^{(i)})\Pi(\mathrm{d}\omega)}{\int_\Omega\prod_{i=1}^n f_\omega(y^{(i)})\Pi(\mathrm{d}\omega)},\ \ A\subseteq\Omega.$$

However, as discussed in the introduction, this posterior distribution is typically intractable. In the frequentist literature, a commonly used approach to circumventing computational difficulties with graphical models consists in replacing the likelihood function by a pseudo-likelihood function. For $\omega \in \mathcal{M}_p$, let $\omega_{\cdot i}$ denote the $i$-th column of $\omega$. Note that in the present case, if $(Y_1, \ldots, Y_p) \sim f_\omega$, then for $1 \leq j \leq p$, the conditional distribution of $Y_j$ given $\{Y_k,\ k \neq j\}$ depends on $\omega$ only through the $j$-th column $\omega_{\cdot j}$. We write this conditional distribution as $u \mapsto f^{(j)}_{\omega_{\cdot j}}(u|y_{-j})$, where for $y \in \mathsf{Y}$, $y_{-j} \stackrel{\text{def}}{=} (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_p)$, (with obvious modifications when $j = 1, p$). Let

$$\tilde{\Omega} \stackrel{\text{def}}{=} \Big\{\omega \in \mathcal{M}_p :\ u \mapsto f^{(j)}_{\omega_{\cdot j}}(u|y_{-j}) \text{ is a well-defined density on } \mathsf{Y}_j, \\ \text{for all } y \in \mathsf{Y},\ \text{ and all } 1 \leq j \leq p\big\}.$$

Note that $\Omega \subseteq \tilde{\Omega}$. The most commonly used pseudo-likelihood method consists in replacing the initial likelihood contribution $f_\omega(y^{(i)})$ by

$$\tilde{f}_\omega(y^{(i)}) = \prod_{j=1}^{p} f^{(j)}_{\omega_{\cdot j}}(y^{(i)}_j | y^{(i)}_{-j}),\ \ \omega \in \tilde{\Omega}. \tag{2}$$

This pseudo-likelihood approach typically brings important simplifications. For instance, in the Gaussian case, the parameter space $\tilde{\Omega}$ corresponds to the space of symmetric matrices with positive diagonals elements, which has a simpler geometry compared to $\mathcal{M}_p^+$. And in the case of discrete graphical models, the conditional models typically have tractable normalizing constants. The idea goes back at least to Besag (1974), and penalized versions of pseudo-likelihood functions have been employed by several authors to fit high-dimensional graphical models. In a Bayesian setting this approach works well for small to moderate size graphs. The issue is that the space $\tilde{\Omega} \subset \mathcal{M}_p$ grows as $O(p^2)$, and MCMC simulation for exploring probability distributions on such very large spaces is inherently a difficult problem (for example, for only $p = 100$ the dimension of $\tilde{\Omega}$ is larger than $1 \times 10^4$).

A related pseudo-likelihood for this problem is suggested by the neighborhood selection of Meinshausen and Buhlmann (2006). The idea consists in relaxing the symmetry constraint in $\tilde{\Omega}$. For $1 \leq j \leq p$, we set

$$\Omega_j \stackrel{\text{def}}{=} \Big\{\theta \in \mathbb{R}^p :\ u \mapsto f^{(j)}_\theta(u|y_{-j}) \text{ is a well-defined density on } \mathsf{Y}_j, \\ \text{for all } y \in \mathsf{Y},\ \text{ and all } 1 \leq j \leq p\big\}.$$

We note that if $\omega \in \Omega$, then $\omega_{\cdot j} \in \Omega_j$. Hence these sets $\Omega_j$ are nonempty, and we define $\check{\Omega} \stackrel{\text{def}}{=} \Omega_1 \times \cdots \times \Omega_p$, that we identify as a subset of the space of $p \times p$ real

matrices $\mathbb{R}^{p\times p}$. In particular if $\omega \in \check{\Omega}$, and consistently with our notation above, $\omega_{\cdot,j}$ denotes the $j$-column of $\omega$. We consider the pseudo-model $\{\check{f}_\omega,\ \omega \in \check{\Omega}\}$, where

$$\check{f}_\omega(y) \stackrel{\text{def}}{=} \prod_{j=1}^{p} f^{(j)}_{\omega_{\cdot,j}}(y_j|y_{-j}),\ \ \omega \in \check{\Omega},\ \ y \in \mathsf{Y}. \tag{3}$$

Notice that by definition $\check{\Omega}$ is a product space, whereas $\tilde{\Omega}$ is not, due to the symmetry constraint. This implies that $\omega \mapsto \check{f}_\omega(y)$ factorizes along the columns of $\omega$, whereas $\omega \mapsto \tilde{f}_\omega(y)$ typically does not. With a prior distribution $\Pi$ on $\check{\Omega}$, the quasi-likelihood function $\omega \mapsto \check{f}_\omega$ leads to a quasi-posterior distribution given by

$$\check{\Pi}_{n,p}(A|x) = \frac{\int_A \prod_{i=1}^n \check{f}_\omega(y^{(i)})\Pi(\mathrm{d}\omega)}{\int_{\check{\Omega}} \prod_{i=1}^n \check{f}_\omega(^{(i)})\Pi(\mathrm{d}\omega)},\ \ A \subset \check{\Omega}.$$

Let us assume that the prior distribution factorizes: $\Pi(\mathrm{d}\omega) = \prod_{j=1}^p \Pi_j(\omega_{\cdot j})$. Then we are led to the quasi-posterior distribution

$$\check{\Pi}_{n,p}(\mathrm{d}u_1, \cdots \mathrm{d}u_p|x) = \prod_{j=1}^{p} \check{\Pi}_{n,p,j}(\mathrm{d}u_j|x), \tag{4}$$

where

$$\check{\Pi}_{n,p,j}(\mathrm{d}u|x) = \frac{\prod_{i=1}^n f^{(j)}_u(y^{(i)}_j|y^{(i)}_{-j})\Pi_j(\mathrm{d}u)}{\int_{\Omega_j} \prod_{i=1}^n f^{(j)}_u(y^{(i)}_j|y^{(i)}_{-j})\Pi_j(\mathrm{d}u)},$$

is a probability measure on $\Omega_j$. Basically, relaxing the symmetry allows us to factorize the quasi-likelihood function and this leads to a factorized quasi-posterior distribution, as in (4). Each component of this quasi-posterior distribution can then be explored independently. Despite its simplicity, when used in a parallel computing environment, this approach increases by one order of magnitude the size of graphical models that can be estimated.

**Remark 3.** The method outlined above bears some distant similarity with variational approximation (Blei et al. (2016)). Variational approximation is a popular numerical approximation technique that approximates a posterior distribution by its best representation within a given family of distributions. In contrast, the approach advocated here is more statistical in nature: we approximate the statistical model by a product of smaller conditional models. In that sense, quasi-models give an approach to strike a better trade-off between robustness and computational tractability on one side, and statistical accuracy on the other.

## 3. Gaussian graphical models

Here we consider the Gaussian case where $\mathsf{Y}_i = \mathbb{R}$, $B_i(x) = x^2/2$, and $B_{ij}(x,y) = xy$. Hence in this case, $\Omega = \mathcal{M}_p^+$, $\tilde{\Omega}$ corresponds to the set of symmetric matrices with positive diagonal elements, and $\check{\Omega}$ is the space of $p \times p$ real matrices (not necessarily symmetric) with positive diagonal. Assuming that the diagonal elements are known and given, we shall identify $\check{\Omega}$ with the matrix space $\mathbb{R}^{(p-1)\times p}$.

If $\vartheta \in \mathcal{M}_p^+$, and $(Y_1, \ldots, Y_p) \sim \mathbf{N}(0, \vartheta^{-1})$, it is well known that for all $j \in \{1, \ldots, p\}$, the conditional distribution of $Y_j$ given all other $Y_k = y_k$, for $k \neq j$ is

$$\mathbf{N}\left( -\sum_{k \neq j} \frac{\vartheta_{kj}}{\vartheta_{jj}} y_k, \frac{1}{\vartheta_{jj}} \right), \tag{5}$$

where $\mathbf{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean $\mu$ and variance $\sigma^2$. Given data $x \in \mathbb{R}^{n\times p}$, given $\sigma_j^2 > 0$, and given these conditional distributions, the product of the quasi-model (3) across the data set gives (up to normalizing constants that we ignore) the quasi-likelihood

$$q(\theta; x) \stackrel{\text{def}}{=} \prod_{j=1}^{p} q_j(\theta_{\cdot j}; x),$$

$$\text{with } \; q_j(\theta_{\cdot j}; x) \stackrel{\text{def}}{=} \exp\left( -\frac{1}{2\sigma_j^2} \|x_{\cdot j} - x^{(j)} \theta_{\cdot j}\|_2^2 \right), \;\; \theta \in \mathbb{R}^{(p-1)\times p}, \tag{6}$$

where $x^{(j)} \in \mathbb{R}^{n\times(p-1)}$ is the matrix obtained from $x$ by removing the $j$-th column, and $x_{\cdot j}$ (resp. $\theta_{\cdot j}$) denotes the $j$-column of $x$ (resp. $\theta$). Given (5), it is clear that $\sigma_j^2$ is a proxy for $1/\vartheta_{jj}$. For the time being, we shall assume that the variance terms $\vartheta_{jj}$ are known, and we will set $\sigma_j^2 = 1/\vartheta_{jj}$. In practice we use an empirical Bayes approach (described below) whereby $\vartheta_{jj}$ is obtained from the data. We combine (6) with a prior distribution $\Pi(\mathrm{d}\theta) = \prod_{j=1}^p \Pi_j(\mathrm{d}\theta_{\cdot j})$ to obtain a quasi-posterior distribution on $\mathbb{R}^{(p-1)\times p}$ given by

$$\check{\Pi}_{n,p}(\mathrm{d}\theta | x) = \prod_{j=1}^{p} \check{\Pi}_{n,p,j}(\mathrm{d}\theta_{\cdot j} | x, \sigma_j^2), \tag{7}$$

where $\check{\Pi}_{n,p,j}(\cdot | x, \sigma_j^2)$ is the probability measure on $\mathbb{R}^{p-1}$ given by

$$\check{\Pi}_{n,p,j}(\mathrm{d}z | x, \sigma_j^2) \propto q_j(z; x) \Pi_j(\mathrm{d}z).$$

Again the main appeal of $\check{\Pi}_{n,p}$ is its factorized form, which implies that Monte Carlo samples from $\check{\Pi}_{n,p}$ can be obtained by sampling in parallel from the $p$ distributions $\check{\Pi}_{n,p,j}$.

3.1. **Prior distribution.** We consider the set up where $p$ is large, and the matrix $\theta \in \mathbb{R}^{(p-1)\times p}$ is sparse. In many applications one is mainly interested in the entries of $\theta$. In such contexts, the use of the separable prior advocated above seems reasonable. For each $j \in \{1, \ldots, p\}$, we build the prior $\Pi_j$ on $\mathbb{R}^{(p-1)}$ as in Castillo et al. (2015). First, let $\Delta_p \stackrel{\text{def}}{=} \{0,1\}^{p-1}$, and let $\{\pi_\delta,\ \delta \in \Delta_p\}$ denote a discrete probability distribution on $\Delta_p$ (which we assume to be the same for all the components $j$). We take $\Pi_j$ as the distribution of the random variable $u \in \mathbb{R}^{p-1}$ obtained as follows.

$\delta_{1:p-1} \stackrel{i.i.d.}{\sim} \mathsf{Ber}(\mathsf{q})$. Given $\delta$, $(u_1, \ldots, u_{p-1})$ are conditionally independent

$$\text{and } u_k|\delta \sim \begin{cases} \mathsf{Dirac}(0) & \text{if } \delta_k = 0 \\ \mathsf{Laplace}\left(\frac{\rho_j}{\sigma_j^2}\right) & \text{if } \delta_k = 1 \end{cases}, \quad (8)$$

where $\mathsf{q} \in (0,1)$ and $\rho_j > 0$ are hyper-parameter, $\sigma_j^2$ is as in (6), $\mathsf{Dirac}(0)$ is the Dirac measure on $\mathbb{R}$ with mass at 0, and for $\rho > 0$, $\mathsf{Laplace}(\rho)$ denotes the Laplace distribution with density $(\rho/2)^{-\rho|x|}$, $x \in \mathbb{R}$.

**Remark 4.** In building the prior for $\theta$, we have ignored some important information, notably the symmetry and the positive definiteness of the true parameter. Therefore our modeling framework may not be appropriate for settings where these aspects of the parameter $\theta$ are of prime interest.

**Remark 5.** When available informative prior can be added by replacing the distribution $\mathsf{Laplace}\left(\frac{\rho_j}{\sigma_j^2}\right)$ by a Laplace distribution of the form $\mathsf{Laplace}\left(\frac{\rho_{jk}}{\sigma_j^2}\right)$, where $\rho_{jk}$ is set to a large value if $\theta_{jk}$ is believed to be small, and conversely, $\rho_{jk}$ is set to a small value if $\theta_{jk}$ is believed to be large. We refer for instance to Greenfield et al. (2013) for an example.

3.2. **Posterior contraction and rate.** We study here the behavior of the posterior distribution given in (7), for large $n, p$ and when the prior is as in (8). We will assume that the observed data matrix $x \in \mathbb{R}^{n\times p}$ is a realization of random matrix $X$ the rows of which are i.i.d. random vectors from a mean-zero Gaussian distribution on $\mathbb{R}^p$ with precision matrix $\vartheta$, with known diagonal elements. More precisely,

**H1.** *For some $\vartheta \in \mathcal{M}_p^+$, $X = Z\vartheta^{-1/2}$, where $Z \in \mathbb{R}^{n\times p}$ is a random matrix with i.i.d. rows drawn from the standard multivariate Gaussian distribution on $\mathbb{R}^p$.*

**Remark 6.** H1 implies that the rows of $X$ are i.i.d. random vectors drawn from the multivariate Gaussian distribution $\mathbf{N}(0, \vartheta^{-1})$. However we expect some robustness of the quasi-posterior distribution with respect to misspecification of the distribution of

$X$. Quasi-models are known to be robust to misspecification of the true data generating distribution (Varin et al. (2011)), provided that that distribution is consistent with the conditional specifications in the quasi-model. In the particular case of the Gaussian graphical model, we also expect our posterior distribution to be somewhat robust to the actual distribution of $Z$ in H1.

From the true precision matrix $\vartheta$, we now derive the true value of the parameter $\theta_\star \in \mathbb{R}^{(p-1)\times p}$ towards which $\check{\Pi}_{n,p}$ is expected to converge. For $j = 1, \ldots, p$, we set

$$\theta_{\star kj} = \begin{cases} -\frac{\vartheta_{kj}}{\vartheta_{jj}}, & \text{for } k = 1, \ldots, j-1, \\ -\frac{\vartheta_{(k+1)j}}{\vartheta_{jj}}, & \text{for } k = j, \ldots, p-1 \end{cases}. \tag{9}$$

Let $\delta_\star \in \{0,1\}^{(p-1)\times p}$ be the sparsity structure of $\theta_\star$, defined as $\delta_{\star kj} = \mathbf{1}_{\{|\theta_{\star kj}|>0\}}$. We set

$$s_{\star j} \stackrel{\text{def}}{=} \sum_{k=1}^{p-1} \mathbf{1}_{\{|\theta_{\star kj}|>0\}}, \quad j = 1, \ldots, p \quad \text{and} \quad s_\star \stackrel{\text{def}}{=} \max_{1\le j\le p} s_{\star j}.$$

Hence $s_{\star j}$ is the degree of node $j$, and $s_\star$ is the maximum node degree in the undirected graph defined by $\vartheta$. The asymptotic behavior of $\check{\Pi}_{n,p}$ depends crucially on certain restricted and $m$-sparse eigenvalues of the true precision matrix $\vartheta$, that we introduce next. We set

$$\underline{\kappa} \stackrel{\text{def}}{=} \inf \left\{ \frac{u'\vartheta u}{\|u\|_2^2} : \ u \in \mathbb{R}^p, \ u \neq 0, \ \text{s.t.} \sum_{k:\ \delta_{\star,k}=0} |u_k| \le 7 \sum_{k:\ \delta_{\star,k}=1} |u_k| \right\}, \tag{10}$$

and for $1 \le s \le p$,

$$\underset{\sim}{\kappa}(s) \stackrel{\text{def}}{=} \inf \left\{ \frac{u'\vartheta u}{\|u\|_2^2} : \ u \in \mathbb{R}^p, \ 1 \le \|u\|_0 \le s \right\},$$

$$\tilde{\kappa}(s) \stackrel{\text{def}}{=} \sup \left\{ \frac{u'\vartheta u}{\|u\|_2^2} : \ u \in \mathbb{R}^p, \ 1 \le \|u\|_0 \le s \right\}. \tag{11}$$

In the above equations, $\|u\|_0$ denotes the number of non-zero components of $u$, and we convene that $\inf \emptyset = +\infty$, and $\sup \emptyset = 0$.

We study the contraction of $\check{\Pi}_{n,p}$ in the norm

$$|\!|\!| \theta |\!|\!| \stackrel{\text{def}}{=} \max_{1\le j\le p} \|\theta_{\cdot j}\|_2.$$

**Theorem 7.** *Assume H1 and (8) with* $\mathsf{q} = \frac{1}{p^{u+1}}$ *for some absolute constant* $u > 0$, *and*

$$\rho_j \stackrel{\text{def}}{=} \max_{1\le k\le p} \|X_{\cdot k}\|_2 \sqrt{\frac{24 \log(p)}{\vartheta_{jj}}}, \qquad 1 \le j \le p. \tag{12}$$

*For $1 \leq j \leq p$, suppose that $\sigma_j^2 = 1/\vartheta_{jj}$, and set*

$$\zeta_j = \frac{4}{u} + s_{\star j} + \frac{2}{u}\left(2 + 6912\frac{\tilde{\kappa}(1)}{\underline{\kappa}} + \frac{1}{24(\log(p))^2}\frac{\tilde{\kappa}(s_\star)}{\tilde{\kappa}(1)}\right) s_{\star j},$$

*$\bar{s}_j \stackrel{\text{def}}{=} \lceil s_{\star j} + \zeta_j \rceil$, and $\bar{s} \stackrel{\text{def}}{=} \max_{1\leq j\leq p} \bar{s}_j$. Then there exist absolute constants $a_0 > 0$, $a_1 > 0$, $a_2 > 0$, $M_0 \geq 2$ such that for all $p \geq a_0$, and*

$$n \geq a_1 \bar{s}\left(1 + \frac{\tilde{\kappa}(1)}{\underline{\kappa}}\right)\log(p), \tag{13}$$

*the following two statements hold:*

$$\mathbb{E}\left[\check{\Pi}_{n,p}\left(\left\{\theta \in \mathbb{R}^{(p-1)\times p} :\ \|\theta_{\cdot j}\|_0 \geq \zeta_j \text{ for some } j\right\}\ |X\right)\right] \leq 2\left(\frac{1}{e^{a_2 n}} + \frac{2}{p}\right), \tag{14}$$

$$\mathbb{E}\left[\check{\Pi}_{n,p}\left(\left\{\theta \in \mathbb{R}^{(p-1)\times p} :\ |\!|\!|\theta - \theta_\star|\!|\!| > M_0\epsilon\right\}|X\right)\right] \leq 3\left(\frac{1}{e^{a_2 n}} + \frac{4}{p}\right). \tag{15}$$

*where $\epsilon > 0$ is given*

$$\epsilon \stackrel{\text{def}}{=} \frac{\sqrt{\tilde{\kappa}(1)}}{\underset{\sim}{\kappa}(\bar{s})}\sqrt{\frac{\bar{s}\log(p)}{n}}.$$

*Proof.* See Section 5.2.1. □

**Remark 8.** Under H1 and the assumed prior, (14) says that for $n, p$ large, if $\theta \sim \check{\Pi}_{n,p}(\cdot|X)$, then with high probability $\|\theta_{\cdot j}\|_0 < \zeta_j$ for all $j \in \{1, \ldots, p\}$. Note that if $\tilde{\kappa}(1)/\underline{\kappa}$ is small – meaning $\vartheta$ is well-conditioned – then $\zeta_j$ is of the same order as $s_{\star,j}$. In other words the main conclusion of (14) is that $\check{\Pi}_{n,p}(\cdot|X)$ concentrates most of its probability mass on matrices that are sparse with a sparsity structure that mirrors that of $\vartheta$, provided that $\vartheta$ is well-conditioned. Furthermore, the well-conditioning of $\vartheta$ is measured in terms of the ratio of the restricted eigenvalues $\tilde{\kappa}(1)/\underline{\kappa}$, not in terms of the ratio of its largest eigenvalue to its smallest eigenvalue.

The behavior of the posterior distribution in practice suggests that the large constant 6912 appearing in the theorem is most likely an artifact of the techniques used in the proof, and can probably be improved.

Equation (15) says that the contraction rate of $\check{\Pi}_{n,p}$ towards $\theta_\star$ in the $|\!|\!|\cdot|\!|\!|$ norm is

$$O\left(\sqrt{\frac{\bar{s}\log(p)}{n}}\right).$$

This result can be used to deduce the rate of convergence in the more standard spectral norm, for easy comparison with existing results. To that end, suppose as in Theorem

7 that the diagonal terms $\vartheta_{jj}$ are known, and let $\theta \sim \check{\Pi}_{n,p}(\cdot|X)$. In view of (9), we can use $\theta$ to approximate $\vartheta$ as follows: define $\tilde{\vartheta} \in \mathbb{R}^{p\times p}$ such that for $1 \leq j \leq p$:

$$\tilde{\vartheta}_{jj} = \vartheta_{jj}, \text{ and } \tilde{\vartheta}_{kj} = \begin{cases} -\vartheta_{jj}\theta_{kj} & \text{for } 1 \leq k \leq j-1 \\ -\vartheta_{jj}\theta_{k-1,j} & \text{for } j+1 \leq k \leq p. \end{cases}$$

Note however that $\tilde{\vartheta}$ is not symmetric in general. Following Sun and Zhang (2013) we symmetrize it, by taking

$$\hat{\vartheta} \stackrel{\text{def}}{=} \underset{V\in\mathbb{R}^{p\times p}:\, V=V'}{\mathsf{Argmin}} \|V - \tilde{\vartheta}\|_1,$$

where $\|A\|_1 \stackrel{\text{def}}{=} \max_j \sum_k |A_{kj}|$, is the matrix operator in 1-norm. The matrix $\hat{\vartheta}$ is not available in closed form, but can be computed by linear programming (Sun and Zhang (2013); Yuan (2010)). For a matrix $A$, let $\|A\|_2$ denote its spectral norm (that is, its matrix operator in 2-norm). We note that for any symmetric matrix $A$,

$$\|A\|_2 \leq \|A\|_1 \leq \max_j \sqrt{\|A_{\cdot j}\|_0}\, |\!|\!| A |\!|\!| .$$

With these notations, and with $\epsilon$ as in Theorem 7, we have:

$$\begin{aligned}\check{\Pi}_{n,p}\left(\left\{\|\hat{\vartheta} - \vartheta\|_2 > M_0\tilde{\kappa}(1)\bar{s}^{1/2}\epsilon\right\}|X\right) &\leq \check{\Pi}_{n,p}\left(\{\|\theta_{\cdot j}\|_0 \geq \zeta_j \text{ for some } j\}\,|X\right) \\ &\quad + \check{\Pi}_{n,p}\left(\left\{\|\hat{\vartheta} - \vartheta\|_2 > M_0\tilde{\kappa}(1)\bar{s}^{1/2}\epsilon, \text{ and } \|\theta_{\cdot j}\|_0 < \zeta_j \text{ for all } j\right\}|X\right).\end{aligned}$$

By (14), the expectation of the first term on the right-hand side of the last display is upper bound by $2\left(\frac{1}{e^{a_2 n}} + \frac{2}{p}\right)$. Whereas for $\|\theta_{\cdot j}\|_0 < \zeta_j$ for all $j$,

$$\|\hat{\vartheta} - \vartheta\|_2 \leq \|\hat{\vartheta} - \vartheta\|_1 \leq \|\tilde{\vartheta} - \vartheta\|_1 \leq \max_{1\leq j\leq p} \vartheta_{jj}\sqrt{\bar{s}_j}\, |\!|\!| \theta - \theta_\star |\!|\!| \leq \bar{s}^{1/2}\tilde{\kappa}(1)\, |\!|\!| \theta - \theta_\star |\!|\!| ,$$

where the second inequality uses the fact that $\hat{\vartheta} - \vartheta$ is the symmetrization of $\tilde{\vartheta} - \vartheta$. It follows from the above and (15) that

$$\mathbb{E}\left[\check{\Pi}_{n,p}\left(\left\{\|\hat{\vartheta} - \vartheta\|_2 > M_0\tilde{\kappa}(1)\bar{s}^{1/2}\epsilon\right\}|X\right)\right] \leq 5\left(\frac{1}{e^{a_2 n}} + \frac{4}{p}\right). \tag{16}$$

Hence, the contraction rate of $\check{\Pi}_{n,p}$ in the spectral norm is

$$O\left(\bar{s}\sqrt{\frac{\log(p)}{n}}\right),$$

which matches the rate of convergence of the frequentist neighborhood selection (Sun and Zhang (2013)). The interesting phenomenon here is that in the high-dimensional

regime with $p$ larger than $n$, this rate is typically better than the rate

$$O\left(\sqrt{\left(\sum_{j=1}^{p} s_{\star j}\right)\frac{\log(p)}{n}}\right),$$

achieved by the full likelihood inference, as derived in the frequentist setting by Rothman et al. (2008), and in the Bayesian setting by Banerjee and Ghosal (2015). These results suggest that in the high-dimensional regime, in addition to their computational convenience, quasi-models are perhaps also statistically more efficient than a full likelihood approach.

## 4. Numerical experiments

4.1. **Fully Bayesian quasi-posterior distribution.** Recall that the prior distribution of $\delta_k$ is $\delta_k \sim \mathsf{Ber}(q)$, with $q = p^{-1-u}$. The posterior distribution $\check{\Pi}_{n,p}$ is fairly robust to the choice of $u$, so throughout the simulations, we set $u = 0.5$. In contrast $\check{\Pi}_{n,p}$ is sensitive to $\rho_j$, so we use a fully Bayesian approach, with a prior distribution $\rho_j \sim \phi$, where $\phi$ is the uniform distribution $\mathbf{U}(a_1, a_2)$ for $a_1 = 10^{-5}$, and $a_2 = 10^5$.

Given $\sigma_j^2$, we obtain a fully specified quasi-posterior distribution

$$\prod_{j=1}^{p} \bar{\Pi}_{n,p,j}(\delta, \mathrm{d}\theta, \mathrm{d}\rho_j | x, \sigma_j^2), \tag{17}$$

where the $j$-th component $\bar{\Pi}_{n,p,j}(\cdot|x,\sigma_j^2)$ can be written as follows. For $\delta \in \Delta_p$, let $\mu_\delta$ be the product measure on $\mathbb{R}^{p-1}$ defined as $\mu_\delta(\mathrm{d}u) = \prod_{j=1}^{p-1} \nu_{\delta_j}(\mathrm{d}u_j)$, where $\nu_0(\mathrm{d}z)$ is the Dirac mass at 0, and $\nu_1(\mathrm{d}z)$ is the Lebesgue measure on $\mathbb{R}$. Then

$$\bar{\Pi}_{n,p,j}(\delta, \mathrm{d}\theta, \mathrm{d}\rho_j | x, \sigma_j^2) \propto q_j(\theta; x) \mathsf{q}^{\|\delta\|_1} (1-\mathsf{q})^{p-\|\delta\|_0-1} \times \left(\frac{\rho_j}{2\sigma_j^2}\right)^{\|\delta\|_1} e^{-\frac{\rho_j}{\sigma_j^2}\|\theta\|_1} \phi(\rho_j) \mu_\delta(\mathrm{d}\theta) \mathrm{d}\rho_j. \tag{18}$$

The quasi-posterior distribution (18) depends on the choice of $\sigma_j^2$. Ideally we would like to set $\sigma_j^2 = 1/\vartheta_{jj}$. However this quantity is unknown. In the simulation we choose $\sigma_j^2$ by empirical Bayes. More precisely, following Reid et al. (2013) we estimate $\sigma_j^2$ by

$$\hat{\sigma}_j^2 = \frac{1}{n - \hat{s}_{\lambda_n}} \left\| x_{\cdot j} - x^{(j)} \hat{\beta}_{\lambda_n} \right\|_2^2, \tag{19}$$

where $\hat{\beta}_\lambda$ is the lasso estimate at regularization level $\lambda$ in the linear regression of $x_{\cdot j}$ (the $j$-th column of $x$) of $x^{(j)}$ (the remaining columns). In the procedure, $\lambda_n$ is selected by 10-fold cross-validation, and $\hat{s}_{\lambda_n}$ is the number of non-zero components of $\hat{\beta}_{\lambda_n}$. We explore this approach in the simulations.

Given $j \in \{1, \ldots, p\}$, sampling from the distribution $\bar{\Pi}_{n,p,j}(\cdot|x)$ given in (18) is a difficult computation task, due to the discrete-continuous mixture prior on $\delta$. Here we follow the approach developed by the author in Atchadé (2015), which produces approximate samples from (18) by sampling from its forward-backward approximation denoted by $\bar{\Pi}^{(\gamma)}_{n,p,j}(\delta, \mathrm{d}\theta, \mathrm{d}\rho_{1j}|x, \sigma_j^2)$ – however other approximations schemes could be used as well (Narisetty and He (2014); Schreck et al. (2013)). The parameter $\gamma \in (0, 1/4]$ controls the quality of the approximation. In all the simulations below, we use $\gamma = 0.2$.

4.2. **Simulation set ups.** Throughout we set the sample size to $n = 250$, and $p \in \{100, 500, 1000\}$. We set $x = Z\vartheta^{-1/2}$, where $Z \in \mathbb{R}^{n\times p}$ has i.i.d. standard Gaussian entries, except in setting (a') where we draw the entries of $Z$ from $\mathbf{U}(-1, 1)$. We consider three settings.

- **(a):** $\vartheta$ is generated as in Setting (c) below, but using $p = 100$ nodes.
- **(a'):** $\vartheta$ is generated as in Setting (c) below, using $p = 100$ nodes, but the entries of $Z$ are drawn from $\mathbf{U}(-1, 1)$.
- **(b):** In this case $p = 500$, and we take $\vartheta$ from the R-package `space` based on the work Peng et al. (2009)[1]. These authors have designed a precision matrix $\vartheta$ that is modular with 5 modules of 100 nodes each. Inside each module, there are 3 hubs with degree around 15, and 97 other nodes with degree at most 4. The total number of edges is 587. The resulting partial correlations fall within $(-0.67, -0.10] \cup [0.10, 0.67)$. As explained in Peng et al. (2009), this type of networks are useful models for biological networks.
- **(c):** In this case $p = 1,000$, and we build $\vartheta$ as follows. First we generate a symmetric sparse matrix $B$ such that the number of off-diagonal non-zeros entries is roughly $2p$. We magnified the signal by adding 3 to all the non-zero entries of $B$ (subtracting 3 for negative non-zero entries). Then we set $\vartheta = B + (\epsilon - \lambda_{\min}(B))I_p$, where $\lambda_{\min}(B)$ is the smallest eigenvalue of $B$, with $\epsilon = 1$. In this example, values of the partial correlations are typically in the range $(-0.46, -0.18] \cup [0.18, 0.48)$.

To evaluate the effect of the hyper-parameter $\sigma_j^2$, we report two sets of results. One where $\sigma_j^2 = 1/\vartheta_{jj}$, and another set of results where $\vartheta_{jj}$ is assumed unknown and we select $\sigma_j^2$ from the data, using the cross-validation estimator described in (19).

In order to mitigate the uncertainty in some of the results reported below, we repeat all the MCMC simulations 20 times. Hence, to summarize, for each setting (a), (b),

[1]The precision matrix used here corresponds to the example "Hub network" in Section 3 of Peng et al. (2009). A non-sparse version of $\vartheta$ is attached to the `space` package

and (c), we generate one precision matrix $\vartheta$. Given $\vartheta$, we generate 20 datasets, and for each dataset, we run two MCMC samplers (one where the $\sigma_j^2$'s are taken as the $1/\vartheta'_{jj}s$, and one where they are estimated from the data).

4.3. **Computation details.** All the simulations were performed on a high-performance computer using 100 cores and Matlab 7.14.

To simulate from $\bar{\Pi}^{(\gamma)}_{n,p,j}(\cdot|x,\sigma_j^2)$ for a given $j$, we run the MCMC sampler for $50,000$ iterations and discard the first $10,000$ iterations as burn-in. From the MCMC output, we estimate the structure $\delta \in \{0,1\}^{p\times p}$ as follows. We set the diagonal of $\delta$ to one, and for each off-diagonal entry $(i,j)$ of $\delta$, we estimate $\delta_{ij}$ as equal to 1 if the sample average estimate of $\delta_{ij}$ (from the $j$-th chain) and the sample average estimate of $\delta_{ji}$ (from the $i$-th chain) are both larger than 0.5. Otherwise $\delta_{ij}=0$. Obviously, other symmetrization rules could be adopted.

Given the estimate $\hat{\delta}$ say, of $\delta$, we estimate $\vartheta \in \mathbb{R}^{p\times p}$ as follows. We set the diagonal of $\vartheta$ to $(1/\sigma_j^2)$. For $i\neq j$, if $\hat{\delta}_{ij}=0$, we set $\vartheta_{ij}=\vartheta_{ji}=0$. Otherwise we estimate $\vartheta_{ij}=\vartheta_{ji}$ as $0.5(-1/\sigma_j^2)\bar{\vartheta}_{ij}+0.5(-1/\sigma_i^2)\bar{\vartheta}_{ji}$, where $\bar{\vartheta}_{ij}$ (resp. $\bar{\vartheta}_{ji}$) is the Monte Carlo sample average estimate of $\vartheta_{ij}$ from the $j$-th chain (resp. $i$-th chain).

For all the off-diagonal components $(i,j)$ such that $\hat{\delta}_{ij}=1$, Bayesian posterior intervals can also be produced by taking the union of the 95% posterior intervals from the $i$-th and $j$-th chains. When $\hat{\delta}_{ij}=0$, those confidence intervals are set to $\{0\}$.

4.4. **Results.** We evaluate the behavior of the quasi-posterior distribution (7) on three simulated datasets. As benchmark, we also report the results obtained using the elastic net estimator

$$\hat{\vartheta}_{\mathsf{glasso}} = \mathsf{Argmin}_{\ \theta\in\mathcal{M}_p^+}\left[-\log\det\theta + \mathsf{Tr}(\theta S) + \lambda\sum_{i,j}\left(\alpha|\theta_{ij}| + \frac{(1-\alpha)}{2}\theta_{ij}^2\right)\right],$$

where $S=(1/n)x'x$, $\alpha=0.9$, and $\lambda>0$ is a regularization parameter. We choose $\lambda$ by minimizing $-\log\det\left(\hat{\theta}(\lambda)\right)+\mathsf{Tr}(\hat{\theta}(\lambda)S)+\log(n)\sum_{i<j}\mathbf{1}_{\{|\hat{\theta}(\lambda)_{ij}|>0\}}$, over a finite set of values of $\lambda$. Our goal is not to compare the quasi-Bayesian method to graphical `lasso`, since the former utilizes vastly more computing power that the latter. Rather, we report these numbers as references that help better understand the behavior of the proposed methodology.

We look at the performance of the method by computing the relative Frobenius norm, the sensitivity and the precision of the estimated matrix (as obtained above).

These quantities are defined respectively as

$$\mathcal{E} = \frac{\|\hat{\vartheta} - \vartheta\|_{\mathsf{F}}}{\|\vartheta\|_{\mathsf{F}}}, \quad \mathsf{SEN} = \frac{\sum_{i<j} \mathbf{1}_{\{|\vartheta_{ij}|>0\}} \mathbf{1}_{\{\mathsf{sign}(\hat{\vartheta}_{ij})=\mathsf{sign}(\vartheta_{ij})\}}}{\sum_{i<j} \mathbf{1}_{\{|\vartheta_{ij}|>0\}}};$$
$$\text{and} \quad \mathsf{PREC} = \frac{\sum_{i<j} \mathbf{1}_{\{|\hat{\vartheta}_{ij}|>0\}} \mathbf{1}_{\{\mathsf{sign}(\hat{\vartheta}_{ij})=\mathsf{sign}(\vartheta_{ij})\}}}{\sum_{i<j} \mathbf{1}_{\{|\hat{\vartheta}_{ij}|>0\}}}. \quad (20)$$

We average these statistics over the 20 simulations replications. We compute also the same quantities for the elastic net $\hat{\vartheta}_{\mathsf{glasso}}$. These results are reported in Table 1-4. These results suggest that the quasi-Bayesian procedure generally has good contraction properties in the Frobenius norm, and the deviation from the Gaussian distribution did affect the procedure in any significant way. The results also suggest that the quasi-Bayesian procedure tends to produce high false-negatives, but has excellent false-positive rates, even with $p = 1,000$. The same conclusion seems to hold across all three network settings considered in the simulations.

We also notice with satisfaction that there seems to be little difference between the results where $\vartheta_{jj}$ is assumed known and the results where $\vartheta_{jj}$ is estimated from the data.

| | $\vartheta_{jj}$ known | Empirical Bayes | `Glasso` |
|---|---|---|---|
| Relative Error ($\mathcal{E}$ in %) | 19.2 | 21.6 | 63.1 |
| Sensitivity (SEN in %) | 68.4 | 69.0 | 40.5 |
| Precision (PREC in %) | 100.0 | 100.0 | 74.9 |

TABLE 1. Table showing the relative error, sensitivity and precision (as defined in (20)) for Setting (a), with $p = 100$ nodes. Based on 20 simulation replications. Each MCMC run is $5 \times 10^4$ iterations.

| | $\vartheta_{jj}$ known | Empirical Bayes | `Glasso` |
|---|---|---|---|
| Relative Error ($\mathcal{E}$ in %) | 18.6 | 21.5 | 61.7 |
| Sensitivity (SEN in %) | 76.4 | 75.6 | 49.8 |
| Precision (PREC in %) | 99.9 | 99.9 | 79.5 |

TABLE 2. Table showing the relative error, sensitivity and precision (as defined in (20)) for Setting (a'), with $p = 100$ nodes. Based on 20 simulation replications. Each MCMC run is $5 \times 10^4$ iterations.

| | $\vartheta_{jj}$ known | Empirical Bayes | `Glasso` |
|---|---|---|---|
| Relative Error ($\mathcal{E}$ in %) | 23.1 | 26.2 | 45.2 |
| Sensitivity (SEN in %) | 44.6 | 45.4 | 87.9 |
| Precision (PREC in %) | 100 | 99.9 | 56.1 |

TABLE 3. Table showing the relative error, sensitivity and precision (as defined in (20)) for Setting (b), with $p = 500$ nodes. Based on 20 simulation replications. Each MCMC run is $5 \times 10^4$ iterations.

| | $\vartheta_{jj}$ known | Empirical Bayes | `Glasso` |
|---|---|---|---|
| Relative Error ($\mathcal{E}$ in %) | 30.8 | 35.2 | 66.9 |
| Sensitivity (SEN in %) | 16.3 | 16.4 | 6.6 |
| Precision (PREC in %) | 99.9 | 99.8 | 94.7 |

TABLE 4. Table showing the relative error, sensitivity and precision (as defined in (20)) for Setting (c), with $p = 1,000$ nodes. Based on 20 simulation replications. Each MCMC run is $5 \times 10^4$ iterations.

4.5. **An illustration with real data.** We illustrate the method with a real data example taken from `http://ccb.nki.nl/data/` and discussed in van de Vijver et al. (2002). Gene expression profiles were obtained for 295 women with breast cancer at the Netherlands Cancer Institute. The group of patients was then followed over time, and we denote by $Y$ the binary variable that is 1 if a metastasis occurs within the first 5 years, and 0 otherwise. Of the group of 295 patients, 101 developed a metastasis within the first 5 years. The use of gene expression profile to identify group of genes that are the most predictive of poor prognosis (metastasis after initial treatment) is an important topic in cancer research. Here to illustrate our methodology, we use a Gaussian graphical model to estimate and compare the gene expression networks for patients with and without distant metastasis.

The initial dataset has 24884 genes. To reduce its size, we perform a logistic regression of $Y$ on each gene, and select only genes for which the p-value of the gene's coefficient is smaller than 0.005. A total of 984 genes was selected. We center and re-scale the gene expressions in each data set, and apply our Bayesian approach as outlined above. All the R and Matlab code are provided in the supplementary material. Figure 1 shows the sparsity structure of the two graphs. The main observation that can be highlighted in Figure 1 is the fact the gene network of the patients with distant metastasis (Group 1) is noticeably more sparse. To help identify the genes whose neighborhood has changed the most, we compute for each gene $j$ the symmetric

difference score

$$S_j = \frac{|\mathcal{N}_{1,j} \triangle \mathcal{N}_{2,j}|}{1 + |\mathcal{N}_{1,j} \cap \mathcal{N}_{2,j}|},$$

where $\mathcal{N}_{i,j}$ is the set of genes that are connected to gene $j$ in network $i$, $A \triangle B$ is the symmetric difference between sets $A$ and $B$, and $|A|$ denotes the number of elements in set $A$. Figure 2 shows the symmetric difference scores, in increasing order. The right-hand side of Figure 2 shows the 20 genes with the highest score, with their names in Table 5.

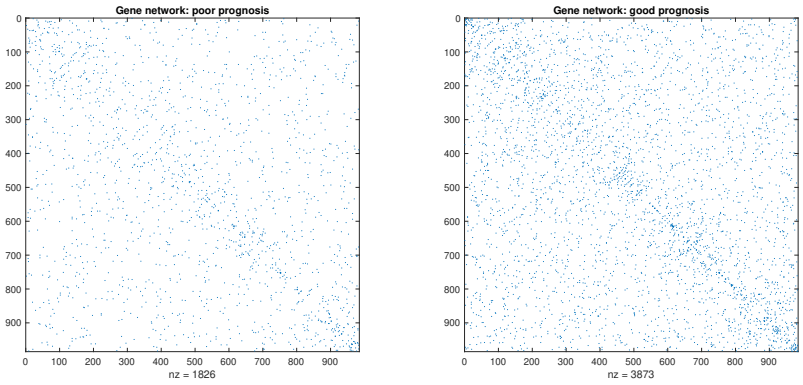


FIGURE 1. Estimated gene networks. Left: from poor prognosis sample, right: from good prognosis samples.

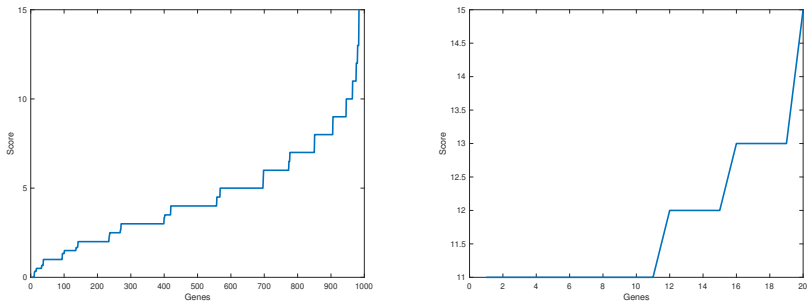


FIGURE 2. Symmetric difference scores of the genes. Right plot shows the 20 genes with the highest score.

| genes (1-10) | genes (11-20) |
|---|---|
| Contig40965_RC | Contig57584_RC |
| X05299 | Contig61227_RC |
| AI632789_RC | Contig8888_RC |
| Contig48471_RC | Contig22253_RC |
| Contig619_RC | Contig51151_RC |
| Contig8156_RC | X71490 |
| NM_001513 | NM_001528 |
| Contig65934_RC | NM_001533 |
| NM_002266 | NM_002269 |
| NM_001540 | Contig17109_RC |

TABLE 5. Names of the 20 genes with the highest symmetric difference scores.

## 5. Proof of Theorem 7

We shall first establish from first principle some contraction properties for posterior distributions in linear regression models. We will then reduce the proof of Theorem 7 to the linear regression case. Our methods of proof are similar to techniques developed in Castillo et al. (2015); Atchade (2017). These ideas extends earlier works on Bayesian asymptotics (see e.g. Ghosal et al. (2000) and the references therein).

5.1. **Posterior contraction of high-dimensional linear regression models.** Let $X \in \mathbb{R}^{n\times p}$ be a design matrix, $\theta_\star \in \mathbb{R}^p$, $\sigma_0^2 > 0$. Suppose that

$$Z \sim \mathbf{N}(X\theta_\star, \sigma_0^2 I_n). \tag{21}$$

We will write $\mathbb{P}_\star$ and $\mathbb{E}_\star$ for the probability measure and expectation operator under the distribution of $Z$ assumed in (21).

Let $\Delta \stackrel{\text{def}}{=} \{0,1\}^p$, and $\{\omega_\delta,\ \delta \in \Delta\}$ be a probability distribution on $\Delta$. For $\sigma^2 > 0$, $\lambda > 0$, we consider the posterior distribution

$$\Pi_n(\mathrm{d}\theta|Z) = \frac{1}{C_n(Z)} \sum_{\delta\in\Delta} \omega_\delta \frac{e^{-\frac{1}{2\sigma^2}\|Z-X\theta\|_2^2}}{e^{-\frac{1}{2\sigma^2}\|Z-X\theta_\star\|_2^2}} \left(\frac{\lambda}{2}\right)^{\|\delta\|_1} e^{-\lambda\|\theta\|_1} \mu_\delta(\mathrm{d}\theta).$$

We make the following assumptions on the distribution $\{\omega_\delta,\ \delta \in \Delta\}$.

**H2.** *For all $\delta \in \Delta$, $\omega_\delta = g_{\|\delta\|_0}\binom{p}{\|\delta\|_0}^{-1}$. Furthermore, there exists universal constants $c_1, c_2, c_3, c_4$ such that for all $s = 1, \ldots, p$,*

$$\left(\frac{c_1}{p^{c_3}}\right) g_{s-1} \le g_s \le \left(\frac{c_2}{p^{c_4}}\right) g_{s-1}\ .$$

**Remark 9.** We note here that if $\delta \stackrel{i.i.d.}{\sim} \mathsf{Ber}(\mathsf{q})$ where $\mathsf{q} = \frac{1}{p^{u+1}}$ as assumed in (8), then $w_\delta = g_{\|\delta\|_0}\binom{p-1}{\|\delta\|_0}^{-1}$, where $g_s = \binom{p}{s}\mathsf{q}^s(1-\mathsf{q})^{p-s-1}$. Furthermore it is easy to check

that $\{g_s\}$ satisfies the double inequalities in H2 with $c_1 = 0.5$, $c_2 = 2$, $c_3 = u+1$, and $c_4 = u$. In other words, the prior distribution chosen in (8) satisfies H2.

Let $\delta_\star \in \Delta$ denote the sparsity structure of $\theta_\star$. That is for all $j \in \{1, \ldots, p\}$, $\delta_{\star,j} = 1$ if and only if $|\theta_{\star,j}| > 0$. Set

$$\mathcal{C} \stackrel{\text{def}}{=} \left\{ \theta \in \mathbb{R}^p : \sum_{j,\, \delta_{\star,j}=1} |\theta_j| \leq 7 \sum_{j,\, \delta_{\star,j}=0} |\theta_j| \right\},$$

and define

$$\underline{v} \stackrel{\text{def}}{=} \inf \left\{ \frac{u'(X'X)u}{n\|u\|_2^2},\ u \neq 0,\ u \in \mathcal{C} \right\}.$$

For integer $s \geq 1$, we define

$$\underline{v}(s) \stackrel{\text{def}}{=} \inf \left\{ \frac{u'(X'X)u}{n\|u\|_2^2},\ u \neq 0,\ \|u\|_0 \leq s \right\}, \quad \bar{v}(s) \stackrel{\text{def}}{=} \sup \left\{ \frac{u'(X'X)u}{n\|u\|_2^2},\ u \neq 0,\ \|u\|_0 \leq s \right\}.$$

**Theorem 10.** *Assume (21) and H2, and suppose that $\underline{v} > 0$. Then there exists a constant $A_0$ that depends only on the constants $c_1, c_2, c_3, c_4$ in H2 such that the following statements hold.*

*(1) For all $p \geq A_0$, and $\zeta > 0$,*

$$\mathbb{E}_\star \left[ \Pi_n \left( \{\theta \in \mathbb{R}^p :\ \|\theta\|_0 \geq s_\star + \zeta\} \,|Z \right) \right] \leq 2 \exp\left( - \frac{\lambda^2 \sigma^4 \log(p)}{8\sigma_0^2 \max_{1 \leq j \leq p} \|X_{\cdot,j}\|_2^2} \right) + 2(4^{s_\star}) e^{\frac{2\lambda^2\sigma^2 s_\star}{n\underline{v}}} \left( 1 + \frac{n\bar{v}(s_\star)}{\lambda^2\sigma^2} \right)^{s_\star} \binom{p}{s_\star} \left( \frac{4c_2}{p^{c_4}} \right)^{\zeta}. \quad (22)$$

*(2) For all $p \geq A_0$, $M \geq 96$, and integer $\bar{s} \geq s_\star$ such that $\underline{v}(\bar{s}) > 0$, set*

$$\underline{\kappa} \stackrel{\text{def}}{=} \frac{n\underline{v}(\bar{s})}{\sigma^2}, \quad \epsilon \stackrel{\text{def}}{=} \frac{2\lambda\sqrt{\bar{s}}}{\underline{\kappa}},$$

*and $\mathcal{A}_\epsilon \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^p :\ \|\theta - \theta_\star\|_0 \leq \bar{s},\ \|\theta - \theta_\star\|_2 \leq M\epsilon\}$. Then*

$$\mathbb{E}_\star \left[ \Pi_n \left( \mathcal{A}_\epsilon | Z \right) \right] \leq 2 \exp\left( - \frac{\lambda^2 \sigma^4 \log(p)}{8\sigma_0^2 \max_{1 \leq j \leq p} \|X_{\cdot,j}\|_2^2} \right) + \binom{p}{\bar{s}} 9^{\bar{s}} \frac{e^{-\underline{\kappa}(M\epsilon)^2/32}}{1 - e^{-\underline{\kappa}(M\epsilon)^2/32}} + 2 \left( 1 + \frac{n\bar{v}(s_\star)}{\sigma^2\lambda^2} \right)^{s_\star} \binom{p}{s_\star} \left( \frac{p^{c_3}}{c_1} \right) \frac{e^{-\underline{\kappa}(M\epsilon)^2/64}}{1 - e^{-\underline{\kappa}(M\epsilon)^2/64}}.$$

*Proof.* We start the proof with some notations. We set

$$\bar{f}_{n,\theta}(z) = \left( \frac{1}{2\pi\sigma} \right)^{n/2} e^{-\frac{1}{2\sigma^2}\|z - X\theta\|_2^2}, \quad z \in \mathbb{R}^n, \quad \theta \in \mathbb{R}^p.$$

$$\begin{aligned}\mathcal{L}_{n,\theta_\star}(\theta;z) &= \log \bar{f}_{n,\theta}(z) - \log \bar{f}_{n,\theta_\star}(z) - \left\langle \nabla \log \bar{f}_{n,\theta_\star}(z), \theta-\theta_\star \right\rangle \\ &= -\frac{n}{2\sigma^2}(\theta-\theta_\star)' \left(\frac{X'X}{n}\right)(\theta-\theta_\star).\end{aligned}$$

We will need the following lemmas which are special cases of respectively Lemma 11 and Lemma 14 of Atchade (2017).

**Lemma 11.** *The normalizing constant of* $\Pi_n$ *satisfies for all* $z \in \mathbb{R}^n$,

$$C_n(z) \geq \omega_{\delta_\star} e^{-\lambda\|\theta_\star\|_1} \left(\frac{\lambda^2}{\lambda^2 + \frac{n\bar{v}(s_\star)}{\sigma^2}}\right)^{s_\star},$$

*where* $s_\star = \|\theta_\star\|_0$.

**Lemma 12** (Existing of test)**.** *Fix* $M \geq 2$, $\bar{s} \geq s_\star$ *an integer, and suppose that* $\underline{v}(\bar{s}) > 0$. *Set*

$$\underline{\kappa} \stackrel{\text{def}}{=} \frac{n\underline{v}(\bar{s})}{\sigma^2}, \quad \epsilon \stackrel{\text{def}}{=} \frac{2\lambda\sqrt{\bar{s}}}{\underline{\kappa}}.$$

*There exists a measurable function* $\phi : \mathbb{R}^n \to [0,1]$ *such that*

$$\mathbb{E}_\star(\phi(Z)) \leq \binom{p}{\bar{s}} 9^{\bar{s}} \frac{e^{-\underline{\kappa}(M\epsilon)^2/32}}{1-e^{-\underline{\kappa}(M\epsilon)^2/32}}.$$

*Furthermore, for any* $\theta \in \mathbb{R}^p$ *such that* $\|\theta-\theta_\star\|_0 \leq \bar{s}$, $\|\theta-\theta_\star\|_2 > jM\epsilon$, *for some* $j \geq 1$,

$$\int_{\mathcal{E}} [1-\phi(z)]\, \bar{f}_{n,\theta}(z)\mathrm{d}z \leq e^{-\underline{\kappa}(jM\epsilon)^2/32}.$$

*Proof of Theorem 10-Part(1).* Set $\mathcal{B} \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^p : \|\theta\|_0 \geq s_\star + \zeta\}$, and

$$\mathcal{E} \stackrel{\text{def}}{=} \left\{ z \in \mathbb{R}^n : \|\nabla \log \bar{f}_{n,\theta_\star}(z)\|_\infty \leq \frac{\lambda}{2} \right\}.$$

We set $\bar{\kappa} = n\bar{v}(s_\star)/\sigma^2$. By Lemma 11, and Fubini's theorem,

$$\begin{aligned}&\mathbb{E}_\star\left[\Pi_n(\mathcal{B}|Z)\right] \leq \mathbb{P}_\star(Z \notin \mathcal{E}) \\ &+ \frac{1}{\omega_{\delta_\star}}\left(1+\frac{\bar{\kappa}}{\lambda^2}\right)^{s_\star} \sum_{\delta:\, \|\delta\|_0 \geq s_\star+\zeta} \omega_\delta \left(\frac{\lambda}{2}\right)^{\|\delta\|_0} \int_{\mathbb{R}^p} \mathbb{E}_\star\left[\frac{\bar{f}_{n,\theta}(Z)}{\bar{f}_{n,\theta_\star}(Z)}\mathbf{1}_{\mathcal{E}}(Z)\right] \frac{e^{-\lambda\|\theta\|_1}}{e^{-\lambda\|\theta_\star\|_1}}\mu_\delta(\mathrm{d}\theta)\end{aligned}$$

The integrand of the integral in the last displayed equation is upper bounded by

$$\Psi(\theta) \stackrel{\text{def}}{=} \exp\left(\frac{\lambda}{2}\|\theta-\theta_\star\|_1 + \lambda\|\theta_\star\|_1 - \lambda\|\theta\|_1\right) \mathbb{E}_\star\left[e^{\mathcal{L}_{n,\theta_\star}(\theta;Z)}\mathbf{1}_{\mathcal{E}}(Z)\right].$$

We have

$$\frac{\lambda}{2}\|\theta-\theta_\star\|_1 + \lambda\|\theta_\star\|_1 - \lambda\|\theta\|_1 \leq -\frac{1}{2}\|\delta_\star^c \cdot (\theta-\theta_\star)\|_1 + \frac{3}{2}\|\delta_\star \cdot (\theta-\theta_\star)\|_1.$$

Hence, if $\theta - \theta_\star \notin \mathcal{C}$, using the concavity of $\mathcal{L}_{n,\theta_\star}$,

$$\Psi(\theta) \leq e^{-\frac{\lambda}{4}\|\theta-\theta_\star\|_1} e^{-\frac{\lambda}{4}\|\delta_\star^c\cdot(\theta-\theta_\star)\|_1 + \frac{7\lambda}{4}\|\delta_\star\cdot(\theta-\theta_\star)\|_1} \leq e^{-\frac{\lambda}{4}\|\theta-\theta_\star\|_1}.$$

However, if $\theta - \theta_\star \in \mathcal{C}$, then

$$\mathbb{E}_\star\left[e^{\mathcal{L}_{n,\theta_\star}(\theta;Z)}\mathbf{1}_{\mathcal{E}}(Z)\right] \leq e^{-\frac{n\underline{v}}{2\sigma^2}\|\theta-\theta_\star\|_2^2},$$

and

$$\Psi(\theta) \leq e^{-\frac{\lambda}{2}\|\theta-\theta_\star\|_1} e^{2\sqrt{s_\star}\lambda\|\theta-\theta_\star\|_2 - \frac{n\underline{v}}{2\sigma^2}\|\theta-\theta_\star\|_2^2} \leq e^{\frac{2\lambda^2 s_\star}{\underline{\kappa}}} e^{-\frac{\lambda}{2}\|\theta-\theta_\star\|_1},$$

where $\underline{\kappa} = n\underline{v}/\sigma^2$. We conclude that

$$\begin{aligned}\mathbb{E}_\star\left[\Pi_n(\mathcal{B}|Z)\right] &\leq \mathbb{P}_\star(Z \notin \mathcal{E}) \\ &\quad + e^{\frac{2\lambda^2 s_\star}{\underline{\kappa}}}\left(1+\frac{\bar{\kappa}}{\lambda^2}\right)^{s_\star}\frac{1}{\omega_{\delta_\star}}\sum_{\delta:\ \|\delta\|_0 \geq s_\star+\zeta}\omega_\delta\left(\frac{\lambda}{2}\right)^{\|\delta\|_0}\int_{\mathbb{R}^p} e^{-\frac{\lambda}{4}\|\theta-\theta_\star\|_1}\mu_\delta(\mathrm{d}\theta), \\ &\leq \mathbb{P}_\star(Z \notin \mathcal{E}) + e^{\frac{2\lambda^2 s_\star}{\underline{\kappa}}}\left(1+\frac{\bar{\kappa}}{\lambda^2}\right)^{s_\star}\frac{1}{\omega_{\delta_\star}}\sum_{\delta:\ \|\delta\|_0 \geq s_\star+\zeta}\omega_\delta 4^{\|\delta\|_0}.\end{aligned}$$

Using H2,

$$\begin{aligned}\frac{1}{\omega_{\delta_\star}}\sum_{\delta:\ \|\delta\|_0 \geq s_\star+\zeta}\omega_\delta 4^{\|\delta\|_0} &= \frac{\binom{p}{s_\star}}{g_{s_\star}}\sum_{j=s_\star+\zeta}^{d} 4^j g_j \leq \frac{\binom{d}{s_\star}}{g_{s_\star}}\sum_{j=s_\star+\zeta}^{p} 4^j\left(\frac{c_2}{p^{c_4}}\right)^{j-s_\star} g_{s_\star} \\ &= \binom{p}{s_\star}4^{s_\star}\sum_{j=s_\star+\zeta}^{d}\left(\frac{4c_2}{p^{c_4}}\right)^{j-s_\star}.\end{aligned}$$

For $p$ large enough so that $\frac{4c_2}{p^{c_4}} < 1$, we have $\sum_{j=s_\star+\zeta}^{d}\left(\frac{4c_2}{p^{c_4}}\right)^{j-s_\star} \leq 2\left(\frac{4c_2}{p^{c_4}}\right)^{\zeta}$. It follows that

$$e^{\frac{2\lambda^2 s_\star}{\underline{\kappa}}}\left(1+\frac{\bar{\kappa}}{\lambda^2}\right)^{s_\star}\frac{1}{\omega_{\delta_\star}}\sum_{\delta:\ \|\delta\|_0 \geq s_\star+\zeta}\omega_\delta 4^{\|\delta\|_0} \leq 2(4^{s_\star})e^{\frac{2\lambda^2 s_\star}{\underline{\kappa}}}\left(1+\frac{\bar{\kappa}}{\lambda^2}\right)^{s_\star}\binom{p}{s_\star}\left(\frac{4c_2}{p^{c_4}}\right)^{\zeta}.$$

It remains only to bound the term $\mathbb{P}_\star(Z \notin \mathcal{E})$. Since $\nabla \log \bar{f}_{n,\theta}(z) = X'(z - X\theta)/\sigma^2$, and since $Z \sim \mathbf{N}(0, \sigma_0^2 I_n)$, standard Gaussian exponential bounds give

$$\mathbb{P}_\star(Z \notin \mathcal{E}) \leq 2p\exp\left(-\frac{\lambda^2\sigma^4}{8\sigma_0^2\max_{1\leq j\leq p}\|X_{\cdot,j}\|_2^2}\right).$$

□

*Proof of Theorem 10-Part(2).* We set

$$\bar{\kappa} \stackrel{\text{def}}{=} \frac{n\bar{v}(s_\star)}{\sigma^2}, \quad \underline{\kappa} \stackrel{\text{def}}{=} \frac{n\underline{v}(\bar{s})}{\sigma^2}, \quad \epsilon \stackrel{\text{def}}{=} \frac{2\lambda\sqrt{\bar{s}}}{\underline{\kappa}}.$$

We also set $\mathcal{A}_\epsilon \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^p : \ \|\theta - \theta_\star\|_0 \leq \bar{s}, \ \|\theta - \theta_\star\|_2 > M\epsilon\}$. We have

$$\begin{aligned}\Pi_n(\mathcal{A}_\epsilon|Z) &\leq 1 - \mathbf{1}_{\mathcal{E}}(Z) + \mathbf{1}_{\mathcal{E}}(Z)\Pi_n(\mathcal{A}_\epsilon|Z) \\ &\leq 1 - \mathbf{1}_{\mathcal{E}}(Z) + \phi(Z) + \mathbf{1}_{\mathcal{E}}(Z)(1-\phi(Z))\Pi_n(\mathcal{A}_\epsilon|Z).\end{aligned}$$

Then by Lemma 11, and Fubini's theorem,

$$\begin{aligned}&\mathbb{E}_\star\left[\Pi_n(\mathcal{A}_\epsilon|Z)\right] \leq \mathbb{P}_\star(Z \notin \mathcal{E}) + \mathbb{E}_\star(\phi(Z)) \\ &\quad + \frac{1}{\omega_{\delta_\star}}\left(1 + \frac{\bar{\kappa}}{\lambda^2}\right)^{s_\star} \sum_{\delta\in\Delta} \omega_\delta \left(\frac{\lambda}{2}\right)^{\|\delta\|_0} \int_{\mathcal{A}_\epsilon} \left[\int_{\mathcal{E}} (1-\phi(z))\, \bar{f}_{n,\theta}(z)\mathrm{d}z\right] \frac{e^{-\lambda\|\theta\|_1}}{e^{-\lambda\|\theta_\star\|_1}}\mu_\delta(\mathrm{d}\theta)\end{aligned}$$

We write $\mathcal{A}_\epsilon = \cup_{j\geq 1}\mathcal{A}_\epsilon(j)$, where

$$\mathcal{A}_\epsilon(j) \stackrel{\text{def}}{=} \left\{\theta \in \mathbb{R}^p : \ \|\theta - \theta_\star\|_0 \leq \bar{s}, \ jM\epsilon < \|\theta - \theta_\star\|_2 \leq (j+1)M\epsilon\right\}.$$

Therefore, and using Lemma 12,

$$\begin{aligned}&\int_{\mathcal{A}_\epsilon} \left[\int_{\mathcal{E}} (1-\phi(z))\, \bar{f}_{n,\theta}(z)\mathrm{d}z\right] \frac{e^{-\lambda\|\theta\|_1}}{e^{-\lambda\|\theta_\star\|_1}}\mu_\delta(\mathrm{d}\theta) \\ &\quad \leq \sum_{j\geq 1} e^{-\frac{\underline{\kappa}}{32}(jM\epsilon)^2} e^{3\lambda\sqrt{\bar{s}}(jM\epsilon)} \int_{\mathcal{A}_\epsilon(j)} e^{-\frac{\lambda}{2}\|\theta-\theta_\star\|_1}\mu_\delta(\mathrm{d}\theta) \leq \left(\frac{4}{\lambda}\right)^{\|\delta\|_0} \sum_{j\geq 1} e^{-\frac{\underline{\kappa}}{64}(jM\epsilon)^2},\end{aligned}$$

given that $M \geq 24$. It is easy to check using H2 that

$$\frac{1}{\omega_{\delta_\star}} \sum_\delta \omega_\delta 2^{\|\delta\|_0} \leq 2\binom{p}{s_\star}\left(\frac{p^{c_3}}{c_1}\right).$$

We can then conclude that

$$\begin{aligned}&\mathbb{E}_\star\left[\Pi_n(\mathcal{A}_\epsilon|Z)\right] \leq \mathbb{P}_\star(Z \notin \mathcal{E}) + \mathbb{E}_\star(\phi(Z)) \\ &\qquad\qquad + 2\left(1 + \frac{\bar{\kappa}}{\lambda^2}\right)^{s_\star} \binom{p}{s_\star}\left(\frac{p^{c_3}}{c_1}\right) \sum_{j\geq 1} e^{-\frac{\underline{\kappa}}{32}(jM\epsilon)^2},\end{aligned}$$

as claimed. □

□

5.2. **Proof of Theorem 7.** We rely on the behavior of some restricted and $m$-sparse eigenvalues concepts that we introduce first. For $z \in \mathbb{R}^{n\times q}$, for some $q \geq 1$, and for $s \geq 1$, we define

$$\underline{\kappa}(s,z) \stackrel{\text{def}}{=} \inf_{\delta\in\{0,1\}^q:\, \|\delta\|_0\leq s} \inf\left\{ \frac{\theta'(z'z)\theta}{n\|\theta\|_2^2} :\ \theta\in\mathbb{R}^q,\ \theta\neq 0,\ \sum_{k:\ \delta_k=0}|\theta_k| \leq 7\sum_{k:\ \delta_k=1}|\theta|\right\},$$

and

$$\underset{\sim}{\kappa}(s,z) \stackrel{\text{def}}{=} \inf\left\{ \frac{\theta'(z'z)\theta}{n\|\theta\|_2^2} :\ \theta\in\mathbb{R}^q,\ 1\leq\|\theta\|_0\leq s\right\},$$

$$\tilde{\kappa}(s,z) \stackrel{\text{def}}{=} \sup\left\{ \frac{\theta'(z'z)\theta}{n\|\theta\|_2^2} :\ \theta\in\mathbb{R}^q,\ 1\leq\|\theta\|_0\leq s\right\}.$$

In the above definition, we convene that $\inf\emptyset = +\infty$, and $\sup\emptyset = 0$. We are interested in the behavior of $\underline{\kappa}(s_\star, X)$, $\underset{\sim}{\kappa}(s, X)$ and $\tilde{\kappa}(s, X)$, when $X$ is the random matrix obtained from assumption H1. We will use the following result taken from Raskutti et al. (2010) Theorem 1, and Rudelson and Zhou (2013) Theorem 3.2, which relates the behavior of $\underline{\kappa}(s_\star, X)$, $\underset{\sim}{\kappa}(s, X)$ and $\tilde{\kappa}(s, X)$ to the corresponding term $\underline{\kappa}$, $\underset{\sim}{\kappa}(s)$ and $\tilde{\kappa}(s)$ of the true precision matrix $\vartheta$ introduced in (10)-(11).

**Lemma 13.** *Assume H1. Then there exists finite universal constant $a_1 > 0$, $a_2 > 0$ such that for the following hold.*

*(1) For all $n \geq a_1 \frac{\tilde{\kappa}(1)}{\underline{\kappa}} s_\star \log(p)$, we have*

$$\mathbb{P}\left[64\underline{\kappa}(s_\star, X) < \underline{\kappa}\right] \leq e^{-a_2 n}.$$

*(2) For integers $1 \leq s \leq p$ and $n \geq a_1 s\log(p)$, we have*

$$\mathbb{P}\left[4\underset{\sim}{\kappa}(s,X) < \underset{\sim}{\kappa}(s) \ \text{ or } \ 4\tilde{\kappa}(s,X) > 9\tilde{\kappa}(s)\right] \leq e^{-a_2 n}.$$

5.2.1. *Proof of Theorem 7-Part(1).* We have

$$\check{\Pi}_{n,p}(\mathrm{d}\theta|X) = \prod_{j=1}^{p} \check{\Pi}_{n,p,j}(\mathrm{d}\theta_{\cdot j}|X),$$

where for $j \in \{1,\ldots,p\}$, $\check{\Pi}_{n,p,j}(\mathrm{d}\theta_{\cdot j}|X)$ is given by

$$\check{\Pi}_{n,p,j}(\mathrm{d}u|X) \propto \sum_{\delta\in\Delta_p} \omega_\delta q_j(u;X) \left(\frac{\rho_j}{2\sigma_j^2}\right)^{\|\delta\|_1} e^{-\frac{\rho_j}{\sigma_j^2}\|u\|_1} \mu_\delta(\mathrm{d}u), \tag{23}$$

and

$$\log q_j(u;X) = -\frac{1}{2\sigma_j^2}\|X_{\cdot j} - X^{(j)}u\|_2^2.$$

For $\ell \geq 1$, we define

$$\mathcal{G}_{n,\ell} \stackrel{\text{def}}{=} \left\{ z \in \mathbb{R}^{n\times \ell} : \ \tilde{\kappa}(s_\star, z) \leq \frac{9}{4}\tilde{\kappa}(s_\star), \ \ \tilde{\kappa}(1,z) \leq \frac{9}{4}\tilde{\kappa}(1), \ \text{ and } \ \underline{\kappa}(s_\star, z) \geq \frac{\underline{\kappa}}{64} \right\}.$$

For any $k_j \geq 0$, we start by noting that

$$\mathbb{E}\left[\check{\Pi}_{n,p}\left(\left\{\theta \in \mathbb{R}^{(p-1)\times p} : \ \|\theta_{\cdot j}\|_0 \geq k_j, \ \text{ for some } j\right\} | X\right)\right] \leq \mathbb{P}(X \notin \mathcal{G}_{n,p}) + \sum_{j=1}^{p} \mathbb{E}\left[\mathbf{1}_{\mathcal{G}_{n,p}}(X)\check{\Pi}_{n,p,j}\left(\mathcal{A}_j | X\right)\right].$$

where $\mathcal{A}_j \stackrel{\text{def}}{=} \{u \in \mathbb{R}^{p-1} : \ \|u\|_0 \geq k_j\}$. We notice that if $X \in \mathcal{G}_{n,p}$, then $X^{(j)} \in \mathcal{G}_{n,p-1}$ for any $1 \leq j \leq p$. We recall that the notation $X^{(j)}$ denotes the matrix obtained by removing the $j$-th column of $X$. Hence

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{\mathcal{G}_{n,p}}(X)\check{\Pi}_{n,p,j}\left(\mathcal{A}_j | X\right)\right] \\ &\quad \leq \mathbb{E}\left[\mathbf{1}_{\mathcal{G}_{n,p-1}}(X^{(j)})\check{\Pi}_{n,p,j}\left(\mathcal{A}_j | X\right)\right] = \mathbb{E}\left[\mathbf{1}_{\mathcal{G}_{n,p-1}}(X^{(j)})\mathbb{E}\left(\check{\Pi}_{n,p,j}\left(\mathcal{A}_j | X\right) | X^{(j)}\right)\right].\end{aligned}$$

We conclude that

$$\mathbb{E}\left[\check{\Pi}_{n,p}\left(\left\{\theta \in \mathbb{R}^{(p-1)\times p} : \ \|\theta_{\cdot j}\|_0 \geq k_j, \ \text{ for some } j\right\} | X\right)\right] \leq \mathbb{P}(X \notin \mathcal{G}_{n,p}) + \sum_{j=1}^{p} \mathbb{E}\left[\mathbf{1}_{\mathcal{G}_{n,p-1}}(X^{(j)})T_j\right], \quad (24)$$

where

$$T_j = \mathbb{E}\left(\check{\Pi}_{n,p,j}\left(\mathcal{A}_j | X\right) | X^{(j)}\right).$$

The key idea of the proof is to notice that $T_j$ is an expected quasi-posterior probability in the linear regression model $X_{\cdot j} = X^{(j)}\beta + \eta$, where $\eta \sim \mathbf{N}(0, (1/\vartheta_{jj})I_n)$. Therefore, by Theorem 10-Part(1), we have

$$\begin{aligned}T_j \leq\ & 2p\exp\left(-\frac{\vartheta_{jj}\rho_j^2}{8\max_{k\neq j}\|X_{\cdot k}\|_2^2}\right) \\ &+ 2(4^{s_{\star j}})\left(1 + \frac{\sigma_j^2 L_j}{\rho_j^2}\right)^{s_{\star j}} e^{\frac{2\rho_j^2 s_{\star j}}{\tau_j \sigma_j^2}} \binom{p-1}{s_{\star j}} \left(\frac{4c_2}{p^{c_4}}\right)^{k_j - s_{\star j}}, \quad (25)\end{aligned}$$

where $L_j = n\tilde{\kappa}(s_\star, X^{(j)})$, and $\tau_j = n\underline{\kappa}(s_\star, X^{(j)})$. Given the choice of $\rho_j$, we see that the first term on the right-hand side of (25) is bounded by

$$2p\exp\left(-3\log(p)\right) = \frac{2}{p^2},$$

Using the fact that for $X^{(j)} \in \mathcal{G}_{n,p-1}$, we have $L_j \leq (9/4)n\tilde{\kappa}(s_\star)$, $\tau_j \geq (1/64)n\underset{\sim}{\kappa}$, it is easy to show that the second term on the right-hand side of (25) is bounded by

$$2\exp\left[s_{\star j}\log(p)\left(\frac{6912}{\sigma_j^2\vartheta_{jj}}\frac{\tilde{\kappa}(s_\star)}{\underset{\sim}{\kappa}}+\frac{\sigma_j^2\vartheta_{jj}}{24(\log(p))^2}\frac{\tilde{\kappa}(s_{\star j})}{\tilde{\kappa}(1)}+\frac{\log(4ep)}{\log(p)}\right)-\frac{c_4}{2}(k_j-s_{\star j})\log(p)\right].$$

With $k_j = \zeta_j$ as given in the statement of the theorem, this latter expression is bounded by $2/(p^2)$. This concludes the proof.

5.2.2. *Proof of Theorem 7-Part(2).* We use the same approach as above. We define $\bar{s}_j = s_{\star j} + \zeta_j$ ($\bar{s}_j = 1$ if $s_{\star j} = 0$), and $\bar{s} = \max_j \bar{s}_j$, and we set

$$\mathcal{G}_{n,q} \stackrel{\text{def}}{=} \left\{z \in \mathbb{R}^{n\times q} :\ \tilde{\kappa}(s_\star, z) \leq \frac{9}{4}\tilde{\kappa}(s_\star),\ \text{and}\ \ \underset{\sim}{\kappa}(\bar{s}, z) \geq \frac{1}{4}\underset{\sim}{\kappa}(\bar{s})\right\}.$$

We also define $\mathcal{U} \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^{(p-1)\times p} :\ \|\theta_{\cdot j} - \theta_{\star\cdot j}\|_2 > \epsilon_j,\ \ \text{for some } j\}$, $\bar{\mathcal{U}} \stackrel{\text{def}}{=} \mathcal{U} \cap \{\theta \in \mathbb{R}^{(p-1)\times p} :\ \|\theta_{\cdot j} - \theta_{\star\cdot j}\|_0 \leq s_{\star j} + \zeta_j\ \ \text{for all } j\}$, and

$$\check{\Pi}_{n,p}(\mathcal{U}|X) \leq \check{\Pi}_{n,p}\left(\{\theta \in \mathbb{R}^{(p-1)\times p} :\ \|\theta_{\cdot j} - \theta_{\star\cdot j}\|_0 > s_{\star j} + \zeta_j\ \ \text{for some } j\}|X\right) + \mathbf{1}_{\mathcal{G}_{n,p}^c}(X) + \mathbf{1}_{\mathcal{G}_{n,p}}(X)\check{\Pi}_{n,p}\left(\bar{\mathcal{U}}|X\right). \quad (26)$$

If for some $j$, $\|\theta_{\cdot j} - \theta_{\star\cdot j}\|_0 > s_{\star j} + \zeta_j$, then we necessarily have $\|\theta_{\cdot j}\|_0 > \zeta_j$. Therefore, by Theorem 7, we have:

$$\mathbb{E}\left[\check{\Pi}_{n,p}\left(\{\theta \in \mathbb{R}^{(p-1)\times p} :\ \|\theta_{\cdot j} - \theta_{\star\cdot j}\|_0 > s_{\star j} + \zeta_j\ \ \text{for some } j\}|X\right)\right] \leq \frac{2}{e^{a_2 n}} + \frac{4}{p}. \quad (27)$$

By Lemma 13, for $n \geq a_1\bar{s}\log(p)$,

$$\mathbb{E}\left[\mathbf{1}_{\mathcal{G}_{n,p}^c}(X)\right] = \mathbb{P}\left[X \notin \mathcal{G}_{n,p}\right] \leq \frac{1}{e^{a_2 n}}. \quad (28)$$

It remains to control the last term on the right-hand side of (26). To do so, we note that if $X \in \mathcal{G}_{n,p}$, then $X^{(j)} \in \mathcal{G}_{n,p-1}$ for all $1 \leq j \leq p$. Hence

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\mathcal{G}_{n,p}}(X)\check{\Pi}_{n,p}\left(\bar{\mathcal{U}}|X\right)\right] &\leq \sum_{j=1}^{p}\mathbb{E}\left[\mathbf{1}_{\mathcal{G}_{n,p-1}}(X^{(j)})\check{\Pi}_{n,p,j}(\mathcal{A}_j|X)\right] \\ &\leq \sum_{j=1}^{p}\mathbb{E}\left[\mathbf{1}_{\mathcal{G}_{n,p-1}}(X^{(j)})\mathbb{E}\left(\check{\Pi}_{n,p,j}(\mathcal{A}_j|X)|X^{(j)}\right)\right], \quad (29)\end{aligned}$$

where $\mathcal{A}_j \stackrel{\text{def}}{=} \{u \in \mathbb{R}^{p-1} :\ \|u - \theta_{\star\cdot j}\|_2 > \epsilon_j,\ \ \text{and}\ \ \|u - \theta_{\star\cdot j}\|_0 \leq \bar{s}_j\}$. As in the proof of Theorem 7, we note that under the conditional distribution of $X_{\cdot j}$ given $X^{(j)}$, the term $\check{\Pi}_{n,p,j}(\mathcal{A}_j|X)$ can be viewed as the posterior distribution in the linear regression model

$X_{\cdot j} = X^{(j)}\beta + \eta$, where $\eta \sim \mathbf{N}(0, (1/\vartheta_{jj})I_n)$. Therefore, using Theorem 10-Part(2), and for any constant $M_0 \geq 96$, we have

$$\mathbb{E}\left(\check{\Pi}_{n,p,j}(\mathcal{A}_j|X)|X^{(j)}\right) \leq 2p\exp\left(-\frac{\vartheta_{jj}\rho_j^2}{8\max_{k\neq j}\|X_{\cdot k}\|_2^2}\right)$$
$$+ e^{\bar{s}_j\log(9p)}\frac{e^{-\frac{M_0^2\tau_j\bar{\epsilon}_j^2}{32}}}{1-e^{-\frac{M_0^2\tau_j\bar{\epsilon}_j^2}{32}}} + 2\binom{p}{s_{\star j}}\left(\frac{p^{c_3}}{c_1}\right)^{s_{\star j}}\left(1+\frac{\sigma_j^2 L_j}{\rho_j^2}\right)^{s_{\star j}}\frac{e^{-\frac{M_0^2\tau_j\bar{\epsilon}_j^2}{64}}}{1-e^{-\frac{M_0^2\tau_j\bar{\epsilon}_j^2}{64}}}, \quad (30)$$

where $\bar{\epsilon}_j = \frac{\rho_j \bar{s}_j^{1/2}}{\tau_j}$, $\tau_j = n\underset{\sim}{\kappa}(\bar{s}_j, X^{(j)})$, and $L_j = n\tilde{\kappa}(s_{\star j}, X^{(j)})$. As seen in the proof of Theorem 7, the first term on the right-hand side of (30) is upper bounded by $2/p^2$.

We have

$$\frac{M_0^2\tau_j\bar{\epsilon}_j^2}{32} \geq \left(\frac{54M_0^2}{32}\frac{1}{\sigma_j^2\vartheta_{jj}}\right)\bar{s}_j\log(p).$$

Hence for $p \geq 24e$, and $\frac{54M_0^2}{32}\frac{1}{\sigma_j^2\vartheta_{jj}} \geq 4$, the second term on the right-hand side of (30) is also upper bounded by $2/p^2$. For $\frac{54M_0^2}{32}\frac{1}{\sigma_j^2\vartheta_{jj}} \geq 4$, the third term is upper bounded by

$$4\exp\left[s_{\star j}\log(p)\left(2+c_3+\frac{\sigma_j^2\vartheta_{jj}}{24(\log(p)^2)}\frac{\tilde{\kappa}(s_{\star j})}{\tilde{\kappa}(1)}\right) - \frac{54M_0^2}{64}\frac{1}{\sigma_j^2\vartheta_{jj}}\bar{s}_j\log(p)\right] \leq \frac{2}{p^2},$$

by choosing $\frac{54M_0^2}{64}\frac{1}{\sigma_j^2\vartheta_{jj}} \geq 2 + \frac{c_4}{2}(2+c_3)$. This concludes the proof.

## Acknowledgements

The author would like to thank Shuheng Zhou for very helpful conversations. This work is partially supported by the NSF, grants DMS 1228164 and DMS 1513040.

## References

Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika* **92** 317–335.

Atchade, Y. A. (2017). On the contraction properties of some high-dimensional quasi-posterior distributions. *Ann. Statist.* **45** 2248–2273.

Atchadé, Y. F. (2015). A Moreau-Yosida approximation scheme for high-dimensional quasi-posterior distributions. *ArXiv e-prints* .

Banerjee, S. and Ghosal, S. (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis* **136** 147 – 162.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2016). Variational Inference: A Review for Statisticians. *ArXiv e-prints* arXiv:1601.00670.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics, Springer, Heidelberg. Methods, theory and applications.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480.

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018.

Dobra, A., Lenkoski, A. and Rodriguez, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.* **106** 1418–1433.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531.

Greenfield, A., Hafemeister, C. and Bonneau, R. (2013). Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics* **29** 1060–1067.
URL `http://dx.doi.org/10.1093/bioinformatics/btt099`

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC.

Khondker, Z. S., Zhu, H., Chu, H., Lin, W. and Ibrahim, J. G. (2013). The Bayesian covariance lasso. *Stat. Interface* **6** 243–259.

Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *J. Comput. Graph. Statist.* **20** 140–157. Supplementary material available online.

Li, C. and Jiang, W. (2014). Model Selection for Likelihood-free Bayesian Methods Based on Moment Conditions: Theory and Numerical Examples. *ArXiv e-prints* .

Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs with the lasso. *Annals of Stat.* **34** 1436–1462.

Mukherjee, S. and Speed, T. P. (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences* **105** 14313–14318.

Narisetty, N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42** 789–817.

Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* **104** 735–746.

Peterson, C., Stingo, F. C. and Vannucci, M. (2015). Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association* **110** 159–174.

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259.

Reid, S., Tibshirani, R. and Friedman, J. (2013). A Study of Error Variance Estimation in Lasso Regression. *ArXiv e-prints* .

Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inf. Theor.* **59** 3434–3447.

Schreck, A., Fort, G., Le Corff, S. and Moulines, E. (2013). A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. *ArXiv e-prints* .

Sun, T. and Zhang, C.-H. (2013). Sparse Matrix Inversion with Scaled Lasso. *Journal of Machine Learning Research* **14** 3385–3418.

van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* **347** 1999–2009.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21** 5–42.
URL `http://www.jstor.org/stable/24309261`

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286.