

# QUASI-BAYESIAN ESTIMATION OF LARGE GAUSSIAN GRAPHICAL MODELS

YVES F. ATCHADÉ

(Apr. 2018; first draft Dec. 2015)

ABSTRACT. This paper deals with the Bayesian estimation of high dimensional Gaussian graphical models. We develop a quasi-Bayesian implementation of the neighborhood selection method of Meinshausen and Bühlmann (2006) for the estimation of large Gaussian graphical models. The method produces a product-form quasi-posterior distribution that can be efficiently explored by parallel computing. We derive a non-asymptotic bound on the contraction rate of the quasi-posterior distribution. The result shows that the proposed quasi-posterior distribution contracts towards the true precision matrix at a rate given by the worst contraction rate of the linear regressions that are involved in the neighborhood selection. We develop a Markov Chain Monte Carlo algorithm for approximate computations, following an approach from Atchadé (2015). We illustrate the methodology with a simulation study.

## 1. INTRODUCTION

We consider the problem of fitting large Gaussian graphical models with diverging number of parameters from limited data. This amounts to estimating a sparse precision matrix  $\vartheta \in \mathcal{M}_p^+$  from  $p$ -dimensional Gaussian observations  $y^{(i)} \in \mathbb{R}^p$ ,  $i = 1, \dots, n$ , 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

---

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.

recent years (Dobra et al. (2011); Lenkoski and Dobra (2011); Khondker et al. (2013); Peterson et al. (2015); Banerjee and Ghosal (2015)). 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 when dealing with large graphical models, some form of approximate inference is inescapable. Building on Atchade (2017), we propose a quasi-Bayesian approach for fitting large Gaussian graphical models. Our general approach to the problem consists in working with a “larger” pseudo-model  $\{\check{f}_\theta, \theta \in \check{\Theta}\}$ , where  $\check{\Theta} \supseteq \mathcal{M}_p^+$ . By pseudo-model we mean that the function  $z \mapsto \check{f}_\theta(z)$  is typically not a density on  $\mathbb{R}^p$ , but  $\check{f}_\theta$  is chosen such that the function  $\theta \mapsto \sum_{i=1}^n \log \check{f}_\theta(y^{(i)})$  is a good candidate for M-estimation of  $\vartheta$ . The enlargement of the model space from  $\mathcal{M}_p^+$  to  $\check{\Theta}$  allows us to relax the positive definiteness constraint. With a prior distribution  $\Pi$  on  $\check{\Theta}$ , we obtain a quasi-posterior distribution denoted  $\check{\Pi}_{n,p}$ .  $\check{\Pi}_{n,p}$  is not a posterior distribution in the strict sense of the term since  $\check{f}_\theta$  is not a proper likelihood function. In the specific case of Gaussian graphical models, we take  $\check{\Theta}$  as the space of matrices with positive diagonals, and take  $z \mapsto \check{f}_\theta(z)$  as the pseudo-model underpinning the neighborhood selection method of Meinshausen and Bühlmann (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 (or quasi-likelihood) function in a Bayesian inference is not new and has been developed in other

contexts, such as in Bayesian semi-parametric inference (Kato (2013); Li and Jiang (2014), and the references therein), in approximate Bayesian computation (Fearnhead and Prangle (2010)). See also Atchade (2017) for more references, and for contraction properties.

We study the contraction properties of the quasi-posterior distribution  $\check{\Pi}_{n,p}$  as  $n, p \rightarrow \infty$ . Under the assumption that there exists a well-conditioned true precision matrix, we show that  $\check{\Pi}_{n,p}$  contracts<sup>1</sup> at the rate  $\sqrt{\frac{\bar{s} \log(p)}{n}}$  (see Theorem 3 for a precise statement), where  $\bar{s}$  can be viewed as an upper-bound on the largest degree in the un-directed graph defined by the true precision matrix. This convergence rate corresponds to the worst convergence rate that one gets from the Bayesian analysis of the linear regressions involved in the neighborhood selection. The condition on the sample size  $n$  for the results mentioned above to hold is  $n = O(\bar{s} \log(p))$ , which shows that the quasi-posterior distribution can concentrate around the true precision matrix, even in cases where  $p$  exceeds  $n$ .

The rest of the paper is organized as follows. Section 2 provides a general discussion of quasi-models and quasi-Bayesian inference. The section ends with the introduction of the proposed quasi-Bayesian distribution, based on the neighborhood selection method of Meinshausen and Bühlmann (2006). 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 simulation study is 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  $i \in \{1, \dots, p\}$ , let  $Y_i$  be a nonempty subset of  $\mathbb{R}$ , and set  $Y \stackrel{\text{def}}{=} Y_1 \times \dots \times Y_p$ , that we assume is equipped with a reference sigma-finite product measure  $dy$ . We first consider a class of Markov random field distributions  $\{f_\omega, \omega \in \Omega\}$  for joint modeling of  $Y$ -valued random variables. Let  $\mathcal{M}_p$  denote the set of all real symmetric  $p \times p$  matrices equipped with the inner product  $\langle A, B \rangle_F \stackrel{\text{def}}{=} \sum_{i \leq j} A_i B_{ij}$ , and norm  $\|A\|_F \stackrel{\text{def}}{=} \sqrt{\langle A, A \rangle_F}$ . As above,  $\mathcal{M}_p^+$  denotes the subset of  $\mathcal{M}_p$  of positive definite matrices. For  $i = 1, \dots, p$ , and  $1 \leq j < k \leq p$ , let  $B_i : Y_i \rightarrow \mathbb{R}$  and  $B_{jk} : Y_j \times Y_k \rightarrow \mathbb{R}$  be non-zero measurable functions that we assume known.

---

<sup>1</sup>The contraction rate is measured in the  $L_{\infty,2}$  matrix norm, defined as the largest  $L_2$  column norms

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

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

We assume that  $\mathcal{Y}$  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  $\mathcal{Y}$  given by

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

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  $\mathcal{Y}$ . If  $(Y_1, \dots, Y_p) \sim f_\omega$ , then the parameter  $\omega$  encodes the conditional independence structure among the  $p$  variables  $(Y_1, \dots, Y_p)$ . 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  $(Y_1, \dots, Y_p)$  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  $\mathcal{Y}_i = \mathbb{R}$ ,  $B_i(x) = x^2/2$ ,  $B_{ij}(x, y) = xy$ ,  $i < j$ . In this case  $\mathcal{Y} = \mathbb{R}^p$  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  $\mathcal{Y}_i = \{1, \dots, M\}$ . In this case,  $\mathcal{Y} = \{1, \dots, M\}^d$  equipped with the counting measure. Since  $\mathcal{Y}$  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  $y^{(1)}, \dots, y^{(n)}$  where  $y^{(i)} = (y_1^{(i)}, \dots, y_p^{(i)})' \in \mathcal{Y}$  is viewed as a column vector. We set  $x \stackrel{\text{def}}{=} [y^{(1)}, \dots, y^{(n)}]' \in \mathbb{R}^{n \times p}$ . Given a prior distribution  $\Pi$  on  $\Omega$ , and given the data  $x$ , the resulting posterior distribution for learning  $\omega$  is

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

However, and as discussed in the introduction, this posterior distribution is typically intractable. In the frequentist literature, a commonly used approach to circumvent 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  $(Y_1, \dots, Y_p) \sim f_\omega$ , 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  $u \mapsto f_{\omega_{\cdot j}}^{(j)}(u|y_{-j})$ , where for  $y \in \mathcal{Y}$ ,  $y_{-j} \stackrel{\text{def}}{=} (y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_p)$ , (with obvious modifications when  $j = 1, p$ ). Let

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

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

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

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 diagonal 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  $\tilde{\Omega} \subset \mathcal{M}_p$  grow 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 Bühlmann (2006). The idea consists in relaxing the symmetry constraint in  $\tilde{\Omega}$ . For  $1 \leq j \leq p$ , we set

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

We note that if  $\omega \in \Omega$ , then  $\omega_{\cdot j} \in \Omega_j$ . Hence these sets  $\Omega_j$  are nonempty, and we define  $\check{\Omega} \stackrel{\text{def}}{=} \Omega_1 \times \dots \times \Omega_p$ , 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

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

One can then maximize a penalized version of  $\omega \mapsto \sum_{i=1}^n \log \check{f}_\omega(y^{(i)})$ , and this corresponds to the neighborhood selection method of (Meinshausen and Bühlmann (2006), see also Sun and Zhang (2013)). 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. One can take advantage of this separability for fast computation if the penalty is also separable. 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

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

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

$$\check{\Pi}_{n,p}(du_1, \dots, du_p|x) = \prod_{j=1}^p \check{\Pi}_{n,p,j}(du_j|x), \quad (4)$$

where

$$\check{\Pi}_{n,p,j}(du|x) = \frac{\prod_{i=1}^n f_u^{(j)}(y_j^{(i)} | y_{-j}^{(i)}) \Pi_j(du)}{\int_{\Omega_j} \prod_{i=1}^n f_u^{(j)}(y_j^{(i)} | y_{-j}^{(i)}) \Pi_j(du)},$$

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.

### 3. GAUSSIAN GRAPHICAL MODELS

Here we consider the Gaussian case where  $\mathcal{Y}_i = \mathbb{R}$ ,  $B_i(x) = x^2/2$ , 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  $(Y_1, \dots, Y_p) \sim f_\vartheta$ , it is well known that for all  $j \in \{1, \dots, p\}$ , the conditional distribution of  $Y_j$  given all other  $Y_k = y_k$ , for  $k \neq j$  is

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

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 (upto normalizing constants that we ignore) the quasi-likelihood

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

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

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  $\Pi(d\theta) = \prod_{j=1}^p \Pi_j(d\theta_{\cdot j})$  to obtain a quasi-posterior distribution on  $\mathbb{R}^{(p-1) \times p}$  given by

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

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

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

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 address here the choice of the prior distribution  $\Pi_j$ . For each  $j \in \{1, \dots, p\}$ , 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.

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

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

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

**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. standard normal entries.

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  $j = 1, \dots, p$ ,  $\theta_{\star kj} = -\vartheta_{kj}/\vartheta_{jj}$ , for  $k = 1, \dots, j-1$ , and  $\theta_{\star kj} = -\vartheta_{(k+1)j}/\vartheta_{jj}$ , for  $k = j, \dots, p-1$ . 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

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

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

$$\underline{\kappa} \stackrel{\text{def}}{=} \inf \left\{ \frac{u' \vartheta u}{\|u\|_2^2} : u \in \mathbb{R}^p, u \neq 0, \text{ s.t. } \sum_{k: \delta_{\star, k} = 0} |u_k| \leq \gamma \sum_{k: \delta_{\star, k} = 1} |u_k| \right\}, \quad (9)$$

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

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

$$\tilde{\kappa}(s) \stackrel{\text{def}}{=} \sup \left\{ \frac{u' \vartheta u}{\|u\|_2^2} : u \in \mathbb{R}^p, 1 \leq \|u\|_0 \leq s \right\}. \quad (10)$$

In the above equations, we convene that  $\inf \emptyset = +\infty$ , and  $\sup \emptyset = 0$ .



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

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

**Theorem 3.** *Assume H1 and (8) with  $q = \frac{1}{p^{u+1}}$  for some absolute constant  $u > 0$ , and*

$$\rho_j \stackrel{\text{def}}{=} \max_{1 \leq k \leq p} \|X_{\cdot k}\|_2 \sqrt{\frac{24 \log(p)}{\vartheta_{jj}}}, \quad 1 \leq j \leq p. \quad (11)$$

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  $\bar{s} \stackrel{\text{def}}{=} \max_{1 \leq j \leq p} \bar{s}_j$ . If  $\min(\underline{\kappa}, \underline{\kappa}(\bar{s})) > 0$ , 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

$$n \geq a_1 \bar{s} \left( 1 + \frac{\tilde{\kappa}(1)}{\underline{\kappa}} \right) \log(p), \quad (12)$$

the following two statements hold:

$$\mathbb{E} \left[ \check{\Pi}_{n,p} \left( \left\{ \theta \in \mathbb{R}^{(p-1) \times p} : \|\theta_{\cdot j}\|_0 \geq \zeta_j \text{ for some } j \right\} \mid X \right) \right] \leq 2 \left( \frac{1}{e^{a_2 n}} + \frac{2}{p} \right), \quad (13)$$

$$\mathbb{E} \left[ \check{\Pi}_{n,p} \left( \left\{ \theta \in \mathbb{R}^{(p-1) \times p} : \|\theta - \theta_{\star}\| > M_0 \epsilon \right\} \mid X \right) \right] \leq 3 \left( \frac{1}{e^{a_2 n}} + \frac{4}{p} \right). \quad (14)$$

where  $\epsilon > 0$  is given

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

*Proof.* See Section 5.2.1. □

**Remark 4.** Under H1 and the assumed prior, (13) says that for  $n, p$  large, if  $\theta \sim \check{\Pi}_{n,p}(\cdot \mid X)$ , then with high probability  $\|\theta_{\cdot j}\|_0 < \zeta_j$  for all  $j \in \{1, \dots, p\}$ . Note that if  $\tilde{\kappa}(1)/\underline{\kappa}$  and  $\frac{1}{(\log(p))^2} (\tilde{\kappa}(s_{\star})/\tilde{\kappa}(1))$  are 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 (13) is that  $\check{\Pi}_{n,p}(\cdot \mid X)$  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. 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.

(14) says that the contraction rate of  $\check{\Pi}_{n,p}$  towards  $\theta_{\star}$  in the  $\|\cdot\|$  norm is  $O(\epsilon)$ . This corresponds to the worst among the rates of contraction of the  $p$  linear regression

problems performed during the neighborhood selection procedure. This rate is similar to the rate of convergence of the (frequentist) neighborhood selection method of Meinshausen and Bühlmann (2006), which is of order

$$\sqrt{\frac{s_\star \log(p)}{n}}, \quad (15)$$

(see the discussion in Section 3.4 of Ravikumar et al. (2011)).

#### 4. NUMERICAL EXPERIMENTS

**4.1. Fully Bayesian quasi-posterior distribution.** Recall that the prior distribution of  $\delta_k$  is  $\delta_k \sim \text{Ber}(q)$ , with  $q = p^{-1-u}$ . The posterior distribution  $\tilde{\Pi}_{n,p}$  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

$$\prod_{j=1}^p \bar{\Pi}_{n,p,j}(\delta, d\theta, d\rho_j | x, \sigma_j^2), \quad (16)$$

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  $\mu_\delta(du) = \prod_{j=1}^{p-1} \nu_{\delta_j}(du_j)$ , where  $\nu_0(dz)$  is the Dirac mass at 0, and  $\nu_1(dz)$  is the Lebesgue measure on  $\mathbb{R}$ . Then

$$\begin{aligned} \bar{\Pi}_{n,p,j}(\delta, d\theta, d\rho_j | x, \sigma_j^2) &\propto q_j(\theta; x) \mathbf{q}^{\|\delta\|_1} (1 - \mathbf{q})^{p - \|\delta\|_0 - 1} \\ &\times \left( \frac{\rho_j}{2\sigma_j^2} \right)^{\|\delta\|_1} e^{-\frac{\rho_j}{\sigma_j^2} \|\theta\|_1} \phi(\rho_j) \mu_\delta(d\theta) d\rho_j. \end{aligned} \quad (17)$$

The quasi-posterior distribution (17) 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

$$\hat{\sigma}_j^2 = \frac{1}{n - \hat{s}_{\lambda_n}} \left\| x_{\cdot j} - x^{(j)} \hat{\beta}_{\lambda_n} \right\|_2^2, \quad (18)$$

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  $j \in \{1, \dots, p\}$ , sampling from the distribution  $\bar{\Pi}_{n,p,j}(\cdot | x)$  given in (17) is a difficult computation task, due the discrete-continuous mixture prior on  $\delta$ . Here we follow the approach developed by the author in Atchadé (2015), which produces

approximate samples from (17) by sampling from its forward-backward approximation denoted  $\tilde{\Pi}_{n,p,j}^{(\gamma)}(\delta, d\theta, d\rho_{1j}|x, \sigma_j^2)$  – 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 simulation below, we use  $\gamma = 0.2$ .

**4.2. Simulation set ups.** We generate a data matrix  $x \in \mathbb{R}^{n \times p}$  with i.i.d. rows from  $\mathbf{N}(0, \vartheta^{-1})$ ,  $\vartheta \in \mathcal{M}_p^+$ . Throughout we set the sample size to  $n = 250$ , and we consider three settings.

- (a):  $\vartheta$  is generated as in Setting (c) below, but using  $p = 100$  nodes.
- (b): In this case  $p = 500$ , and we take  $\vartheta$  from the R-package `space` based on the work Peng et al. (2009)<sup>2</sup>. 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-zeros 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 (18).

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), 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).

---

<sup>2</sup>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

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

To simulate from  $\bar{\Pi}_{n,p,j}^{(\gamma)}(\cdot|x, \sigma_j^2)$  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

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

where  $S = (1/n)x'x$ ,  $\alpha = 0.9$ , and  $\lambda > 0$  is a regularization parameter. We choose  $\lambda$  by minimizing  $-\log \det(\hat{\theta}(\lambda)) + \text{Tr}(\hat{\theta}(\lambda)S) + \log(n) \sum_{i < j} \mathbf{1}_{\{|\hat{\theta}(\lambda)_{ij}| > 0\}}$ , 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 than 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

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

$$\text{and } \text{PREC} = \frac{\sum_{i < j} \mathbf{1}_{\{|\hat{\vartheta}_{ij}| > 0\}} \mathbf{1}_{\{\text{sign}(\hat{\vartheta}_{ij}) = \text{sign}(\vartheta_{ij})\}}}{\sum_{i < j} \mathbf{1}_{\{|\hat{\vartheta}_{ij}| > 0\}}}. \quad (19)$$

We average these statistics over the 20 simulations replications. We compute also the same quantities for the elastic net  $\hat{\vartheta}_{\text{glasso}}$ . These results are reported in Table 1-3. These results suggest that the quasi-Bayesian procedure generally has good contraction properties in the Frobenius norm (and hence in the  $L_{\infty,2}$  norm). 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.

	$\vartheta_{jj}^2$ 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 (19)) for Setting (a), with  $p = 100$  nodes. Based on 20 simulation replications. Each MCMC run is  $5 \times 10^4$  iterations.

	$\vartheta_{jj}^2$ 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 2. Table showing the relative error, sensitivity and precision (as defined in (19)) for Setting (b), with  $p = 500$  nodes. Based on 20 simulation replications. Each MCMC run is  $5 \times 10^4$  iterations.

	$\vartheta_{jj}^2$ 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 3. Table showing the relative error, sensitivity and precision (as defined in (19)) for Setting (c), with  $p = 1,000$  nodes. Based on 20 simulation replications. Each MCMC run is  $5 \times 10^4$  iterations.

## 5. PROOF OF THEOREM 3

First we shall establish from first principle some contraction properties for posterior distributions in linear regression models. We will then reduce the proof of Theorem 3 to the linear regression case.

**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

$$Z \sim \mathbf{N}(X\theta_\star, \sigma_0^2 I_n). \quad (20)$$

We will write  $\mathbb{P}_\star$  and  $\mathbb{E}_\star$  for the probability measure and expectation operator under the distribution of  $Z$  assumed in (20).

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

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

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  $s = 1, \dots, p$ ,

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

**Remark 5.** We note here that if  $\delta \stackrel{i.i.d.}{\sim} \text{Ber}(\mathbf{q})$  where  $\mathbf{q} = \frac{1}{p^{u+1}}$  as assumed in (8), then  $w_\delta = g_{\|\delta\|_0} \binom{p-1}{\|\delta\|_0}^{-1}$ , where  $g_s = \binom{p}{s} \mathbf{q}^s (1-\mathbf{q})^{p-s-1}$ . 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  $j \in \{1, \dots, p\}$ ,  $\delta_{\star,j} = 1$  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 6.** *Assume (20) 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  $\zeta > 0$ ,*

$$\begin{aligned} \mathbb{E}_\star [\Pi_n (\{\theta \in \mathbb{R}^p : \|\theta\|_0 \geq s_\star + \zeta\} | Z)] &\leq 2 \exp \left( -\frac{\lambda^2 \sigma^4 \log(p)}{8\sigma_0^2 \max_{1 \leq j \leq p} \|X_{\cdot, j}\|_2^2} \right) \\ &\quad + 2(4^{s_\star}) e^{\frac{2\lambda^2 \sigma^2 s_\star}{n\underline{v}}} \left( 1 + \frac{n\bar{v}(s_\star)}{\lambda^2 \sigma^2} \right)^{s_\star} \binom{p}{s_\star} \left( \frac{4c_2}{p^{c_4}} \right)^\zeta. \end{aligned} \quad (21)$$

(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

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

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

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

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

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

**Lemma 7.** *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 8** (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  $\phi : \mathbb{R}^n \rightarrow [0, 1]$  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  $j \geq 1$ ,

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

*Proof of Theorem 6-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 7, and Fubini's theorem,

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

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

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

and

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



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

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

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 6-Part(2).* We set

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

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 7, and Fubini's theorem,

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

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

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

Therefore, and using Lemma 8,

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

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

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

as claimed.  $\square$

$\square$

$\square$

**5.2. Proof of Theorem 3.** 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

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

In the above definition, we convene that  $\inf \emptyset = +\infty$ , and  $\sup \emptyset = 0$ . We are interested in the behavior of  $\underline{\kappa}(s_\star, X)$ ,  $\underline{\kappa}(s, X)$  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  $\underline{\kappa}(s_\star, X)$ ,  $\underline{\kappa}(s, X)$  and  $\tilde{\kappa}(s, X)$  to the corresponding term  $\underline{\kappa}$ ,  $\underline{\kappa}(s)$  and  $\tilde{\kappa}(s)$  of the true precision matrix  $\vartheta$  introduced in (9)-(10).

**Lemma 9.** *Assume H1. Then there exists finite universal constant  $a_1 > 0$ ,  $a_2 > 0$  such that for the following hold.*

(1) *If  $\underline{\kappa} > 0$ , then for all  $n \geq a_1 \frac{\tilde{\kappa}(1)}{\underline{\kappa}} s_\star \log(p)$*

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

(2) *Let  $1 \leq s \leq p$  be such that  $\underline{\kappa}(s) > 0$ , then for all  $n \geq a_1 s \log(p)$ ,*

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

5.2.1. *Proof of Theorem 3-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  $j \in \{1, \dots, p\}$ ,  $\check{\Pi}_{n,p,j}(\mathrm{d}\theta_{\cdot j}|X)$  is given by

$$\check{\Pi}_{n,p,j}(\mathrm{d}u|X) \propto q_j(u; X) \sum_{\delta \in \Delta_p} \pi_\delta \left( \frac{\rho_j}{2\sigma_j^2} \right)^{\|\delta\|_1} e^{-\frac{\rho_j}{\sigma_j^2} \|u\|_1} \mu_\delta(\mathrm{d}u), \quad (22)$$

and

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

For  $q \geq 1$ , we define

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

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

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

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  $j$  column of  $X$ . Hence

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

We conclude that

$$\begin{aligned} & \mathbb{E} \left[ \check{\Pi}_{n,p} \left( \left\{ \theta \in \mathbb{R}^{(p-1) \times p} : \|\theta_{\cdot j}\|_0 \geq k_j, \text{ for some } j \right\} | X \right) \right] \\ & \leq \mathbb{P}(X \notin \mathcal{G}_{n,p}) + \sum_{j=1}^p \mathbb{E} \left[ \mathbf{1}_{\mathcal{G}_{n,p-1}}(X^{(j)}) T_j \right], \quad (23) \end{aligned}$$

where

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

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 6-Part(1), we have

$$\begin{aligned} T_j & \leq 2p \exp \left( -\frac{\vartheta_{jj}\rho_j^2}{8 \max_{k \neq j} \|X_{\cdot k}\|_2^2} \right) \\ & \quad + 2(4^{s_{\star j}}) \left( 1 + \frac{\sigma_j^2 L_j}{\rho_j^2} \right)^{s_{\star j}} e^{\frac{2\rho_j^2 s_{\star j}}{\tau_j \sigma_j^2}} \binom{p-1}{s_{\star j}} \left( \frac{4c_2}{p^{c_4}} \right)^{k_j - s_{\star j}}, \quad (24) \end{aligned}$$

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 (24) is bounded by

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

Using the fact that for  $X^{(j)} \in \mathcal{G}_{n,p-1}$ , we have  $L_j \leq (9/4)n\tilde{\kappa}(s_{\star})$ ,  $\tau_j \geq (1/64)n\underline{\kappa}$ , it is easy to show that the second term on the right-hand side of (24) is bounded by

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

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 3-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

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

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

$$\begin{aligned} \check{\Pi}_{n,p}(\mathcal{U}|X) &\leq \check{\Pi}_{n,p}(\{\theta \in \mathbb{R}^{(p-1) \times p} : \|\theta_{\cdot j} - \theta_{\star j}\|_0 > s_{\star j} + \zeta_j \text{ for some } j\}|X) \\ &\quad + \mathbf{1}_{\mathcal{G}_{n,p}^c}(X) + \mathbf{1}_{\mathcal{G}_{n,p}}(X) \check{\Pi}_{n,p}(\bar{\mathcal{U}}|X). \end{aligned} \quad (25)$$

If for some  $j$ ,  $\|\theta_{\cdot j} - \theta_{\star j}\|_0 > s_{\star j} + \zeta_j$ , then we necessarily have  $\|\theta_{\cdot j}\|_0 > \zeta_j$ . Therefore, by Theorem 3, we have:

$$\mathbb{E} \left[ \check{\Pi}_{n,p}(\{\theta \in \mathbb{R}^{(p-1) \times p} : \|\theta_{\cdot j} - \theta_{\star j}\|_0 > s_{\star j} + \zeta_j \text{ for some } j\}|X) \right] \leq \frac{2}{e^{a_2 n}} + \frac{4}{p}. \quad (26)$$

By Lemma 9, for  $n \geq a_1 \bar{s} \log(p)$ ,

$$\mathbb{E} \left[ \mathbf{1}_{\mathcal{G}_{n,p}^c}(X) \right] = \mathbb{P}[X \notin \mathcal{G}_{n,p}] \leq \frac{1}{e^{a_2 n}}. \quad (27)$$

It remains to control the last term on the right-hand side of (25). 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

$$\begin{aligned} \mathbb{E} \left[ \mathbf{1}_{\mathcal{G}_{n,p}}(X) \check{\Pi}_{n,p}(\bar{\mathcal{U}}|X) \right] &\leq \sum_{j=1}^p \mathbb{E} \left[ \mathbf{1}_{\mathcal{G}_{n,p-1}}(X^{(j)}) \check{\Pi}_{n,p,j}(\mathcal{A}_j|X) \right] \\ &\leq \sum_{j=1}^p \mathbb{E} \left[ \mathbf{1}_{\mathcal{G}_{n,p-1}}(X^{(j)}) \mathbb{E} \left( \check{\Pi}_{n,p,j}(\mathcal{A}_j|X) | X^{(j)} \right) \right], \end{aligned} \quad (28)$$

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 3, 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 6-Part(2), and for any constant  $M_0 \geq 96$ , we have

$$\begin{aligned} \mathbb{E} \left( \check{\Pi}_{n,p,j}(\mathcal{A}_j|X) | X^{(j)} \right) &\leq 2p \exp \left( -\frac{\vartheta_{jj} \rho_j^2}{8 \max_{k \neq j} \|X_{\cdot k}\|_2^2} \right) \\ &\quad + e^{\bar{s}_j \log(9p)} \frac{e^{-\frac{M_0^2 \tau_j \bar{\epsilon}_j^2}{32}}}{1 - e^{-\frac{M_0^2 \tau_j \bar{\epsilon}_j^2}{32}}} + 2 \binom{p}{s_{\star j}} \left( \frac{p^{c_3}}{c_1} \right)^{s_{\star j}} \left( 1 + \frac{\sigma_j^2 L_j}{\rho_j^2} \right)^{s_{\star j}} \frac{e^{-\frac{M_0^2 \tau_j \bar{\epsilon}_j^2}{64}}}{1 - e^{-\frac{M_0^2 \tau_j \bar{\epsilon}_j^2}{64}}}, \end{aligned} \quad (29)$$

where  $\bar{\epsilon}_j = \frac{\rho_j \bar{s}_j^{1/2}}{\tau_j}$ ,  $\tau_j = n\kappa(\bar{s}_j, X^{(j)})$ , and  $L_j = n\tilde{\kappa}(s_{\star j}, X^{(j)})$ . As seen in the proof of Theorem 3, the first term on the right-hand side of (29) 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 (29) 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.
- ATCHADÉ, 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.
- 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.
- FEARNHEAD, P. and PRANGLE, D. (2010). Constructing summary statistics for approximate bayesian computation: Semi-automatic abc. *Technical Report, Lancaster University, UK* .
- HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC.
- KATO, K. (2013). Quasi-Bayesian analysis of nonparametric instrumental variables models. *Ann. Statist.* **41** 2359–2390.
- 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.
- NARISSETTY, 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.

- RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing  $\ell_1$ -penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980.
- REID, S., TIBSHIRANI, R. and FRIEDMAN, J. (2013). A Study of Error Variance Estimation in Lasso Regression. *ArXiv e-prints* .
- 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.