Bayesian Estimation of Gaussian Conditional Random Fields

Lingrui Gan, Naveen N. Narisetty, and Feng Liang

Department of Statistics, University of Illinois at Urbana-Champaign

Supplementary Material

Section S1 provides a proof of Proposition 1 stated in the main paper.

Section S2 presents useful calculations related to the log-likelihood function.

Section S3 provides proofs for all the technical results from the main paper.

Section S4 provides details of our proposed EM algorithm and its derivation.

Section S5 provides additional simulation results omitted in the main paper.

S1 The spike-and-slab Lasso penalty $\mathrm{pen}_{\mathrm{SS}}(\cdot)$

In this section, we present the proof of Proposition 1 stated in the main paper.

Proof. Throughout the proof, assume $\theta > 0$. Recall $\mathrm{pen}_{\mathrm{SS}}(\theta) = \log \pi(\theta)$.

We have

$$
\begin{aligned}
\text{pen}'_{\text{SS}}(\theta) = -\frac{\pi'(\theta)}{\pi(\theta)} &= \frac{\frac{1}{v_1}\frac{\eta}{2v_1}\exp\big\{-\frac{|\theta|}{v_1}\big\} + \frac{1}{v_0}\frac{(1-\eta)}{2v_0}\exp\big\{-\frac{|\theta|}{v_0}\big\}}{\pi(\theta)} \\
&= \frac{1}{v_1}\eta(\theta) + \frac{1}{v_0}\big(1-\eta(\theta)\big) \\
&= \Lambda(\theta)
\end{aligned}
$$

where $\Lambda(\theta)$ is a discrete random variable:

$$
\Lambda(\theta) = \begin{cases} \frac{1}{v_1}, & \text{w/p } \eta(\theta); \\ \frac{1}{v_0}, & \text{w/p } 1-\eta(\theta); \end{cases}
$$

and $0 \le \eta(\theta) \le 1$ is the conditional probability of θ belonging to the "slab" given that θ is from a mixture of "slab" and "spike", namely,

$$
\begin{aligned}
\eta(\theta) &= \frac{\frac{\eta}{2v_1}\exp\big\{-\frac{|\theta|}{v_1}\big\}}{\frac{\eta}{2v_1}\exp\big\{-\frac{|\theta|}{v_1}\big\} + \frac{1}{v_0}\frac{(1-\eta)}{2v_0}\exp\big\{-\frac{|\theta|}{v_0}\big\}} \\
&= \frac{1}{1+\big(\frac{1-\eta}{\eta}\big)\big(\frac{v_1}{v_0}\big)\exp\big\{-|\theta|(\frac{1}{v_0}-\frac{1}{v_0})\big\}}.
\end{aligned}
$$

Next we compute the second derivative:

$$
\begin{aligned}
\text{pen}''_{\text{SS}}(\theta) &= -\frac{\pi''(\theta)\pi(\theta) - [\pi'(\theta)]^2}{\pi^2(\theta)} \\
&= \Big[\frac{\pi'(\theta)}{\pi(\theta)}\Big]^2 - \frac{\pi''(\theta)}{\pi(\theta)} \\
&= \big[\Lambda(\theta)\big]^2 - \big[\Lambda(\theta)^2\big] = -\text{Var}(\Lambda(\theta)).
\end{aligned}
$$

Since $\text{pen}''_{\text{SS}}(\theta) = -\text{Var}(\Lambda(\theta)) \le 0$. $\text{pen}_{\text{SS}}(\theta)$ is concave. □

S2 The log-likelihood function

Below we list useful calculations related to the log-likelihood function

$$\ell(\Phi) = \frac{n}{2}\Big(\log\det(\Lambda) - \mathrm{tr}(S_{yy}\Lambda + 2S_{xy}\Theta^T + \Lambda^{-1}\Theta^T S_{xx}\Theta)\Big). \tag{S2.1}$$

The gradient and Hessian matrix of the log-likelihood function $\ell(\Phi)$ take the following forms:

$$\nabla\ell(\Phi) = \begin{bmatrix} \nabla_\Lambda \ell(\Phi) \\ \nabla_\Theta \ell(\Phi) \end{bmatrix}, \qquad \nabla^2\ell(\Phi) = \begin{bmatrix} \nabla^2_{\Lambda,\Lambda}\ell(\Phi) & \nabla^2_{\Lambda,\Theta}\ell(\Phi) \\ \nabla^2_{\Lambda,\Theta}\ell(\Phi) & \nabla^2_{\Theta,\Theta}\ell(\Phi) \end{bmatrix},$$

where

$$\begin{aligned} \nabla_\Lambda \ell(\Phi) &= \frac{n}{2}\Big(- S_{yy} + \Lambda^{-1} + \Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1}\Big) \\ \nabla_\Theta \ell(\Phi) &= \frac{n}{2}\Big(- 2S_{xy} - 2S_{xx}\Theta\Lambda^{-1}\Big), \end{aligned} \tag{S2.2}$$

and

$$\nabla^2\ell(\Phi) = \frac{n}{2}\begin{bmatrix} -\Lambda^{-1} \otimes (\Lambda^{-1} + 2\Lambda^{-1}\Theta^T S_{xx}\Theta\Lambda^{-1}) & 2\Lambda^{-1}\otimes\Lambda^{-1}\Theta^T S_{xx} \\ 2\Lambda^{-1}\otimes\Lambda^{-1}\Theta^T S_{xx} & -2\Lambda^{-1}\otimes S_{xx} \end{bmatrix}. \tag{S2.3}$$

Throughout we use $\nabla\ell(\tilde{\Phi})$ to denote the derivative evaluated at $\tilde{\Phi}$. For example, $\nabla\ell(\Phi^0)$ denotes the derivative evaluated at the true parameter value Φ^0 and $H := \nabla^2\ell(\Phi^0)$ denotes the Hessian matrix (S2.3) evaluated at Φ^0.

S3 Proofs of main theorems

S3.1 Proof of Theorem 1

Proof. By Lemma 1, we have

$$\gamma_n := \|\nabla \ell(\Phi^0)\|_{\max} \leq K_* \sqrt{n \log(10(p+q)^2/\eta)} \sim \sqrt{n \log(p+q)}.$$

By Lemma 2, we have have $\beta(\Phi^0; r, \alpha) \geq n\beta_0$ for $r \leq r_0$.

Let $\Delta = \Phi - \Phi^0$. We first prove the result by assuming that $\|\Delta\|_F^2 \leq r_0$ and then show that $\|\Delta\|_F^2 \leq r_0$ must hold under our specified conditions.

If $\|\Delta\|_F^2 \leq r_0$, due to the local strong convexity, we have

$$\begin{aligned}
0 &\geq \ell(\Phi^0 + \Delta) - \ell(\Phi^0) + \text{Pen}(\Phi^0 + \Delta) - \text{Pen}(\Phi^0) \\
&\geq \langle \nabla \ell(\Phi^0), \Delta \rangle + n\beta_0 \|\Delta\|_F^2 + (1/v_1)\|\Delta_{S_0^c}\|_1 - (1/v_0)\|\Delta_{S_0}\|_1 \\
&\geq n\beta_0 \|\Delta\|_F^2 + (1/v_1 - \gamma_n)\|\Delta_{S_0^c}\|_1 - (1/v_0 + \gamma_n)\|\Delta_{S_0}\|_1 \\
&\geq n\beta_0 \|\Delta\|_F^2 - (1/v_0 + \gamma_n)\|\Delta_{S_0}\|_1,
\end{aligned} \tag{S3.4}$$

where the last inequality holds since $1/v_1 > \gamma_n$. Re-arranging the right hand side above and using $\|\Delta_{S_0}\|_1 \leq \sqrt{|S_0|}\|\Delta_{S_0}\|_F \leq \sqrt{|S_0|}\|\tilde{\Delta}\|_F$, we have

$$\|\Delta\|_F \leq \frac{\gamma_n + 1/v_0}{n\beta_0}\sqrt{|S_0|} \leq \left(\frac{1}{nv_1} + \frac{1}{nv_0}\right)\frac{\sqrt{|S_0|}}{\beta_0} = \varepsilon_n \tag{S3.5}$$

where we use the fact $\gamma_n + 1/v_0 \leq 1/v_1 + 1/v_0 = \alpha/v_1$.

If $\|\Delta\|_F^2 > r_0$, as shown in the proof for Proposition 3 in **?**, inequalities in (S3.4) still hold except that β_0 is replaced by

$$b_n = 2\beta_0 \frac{\lambda_{\max}(\Lambda^0)}{\lambda_{\max}(\Lambda)} \geq 2\beta_0 \frac{\lambda_{\max}(\Lambda^0)}{R}$$

where the last inequality is due to the side constraint on Λ: $\lambda_{\max}(\Lambda) \leq R$. Then similar to (S3.5), we have

$$\|\Delta\|_F \leq \frac{\gamma_n + 1/v_0}{nb_n}\sqrt{|S_0|} \leq \frac{R}{2\lambda_{\max}(\Lambda^0)}\varepsilon_n < \sqrt{r_0}$$

by our choice of R, which contradicts with the assumption that $\|\Delta\|_F^2 > r_0$. $\square$

S3.2 Proof of Theorem 2

Proof. First note that with the choice of prior parameters v_1, v_0, η, we have

$$\frac{\eta(\delta_0)}{1-\eta(\delta_0)} = \frac{\eta}{1-\eta}\frac{v_0}{v_1}\exp\left(\delta_0(\frac{1}{v_0} - \frac{1}{v_1})\right) \to \infty,$$

there $\eta(\delta_0) \to 1$ and there exists some constant $\mu > 0$ such that $\mathrm{pen}''_{\mathrm{SS}}(\theta) \geq -\mu/2$ for any $|\theta| > \delta_0$ where $\delta_0 > 0$ is any small positive constant.

Then we prove Theorem 2 with the following three steps:

- *Step 1:* Consider a restricted optimization problem:

$$\min_{\Lambda \succ 0 \Phi_{S_0^c}=0} L(\Phi), \tag{S3.6}$$

- *Step 2:* Show the solution set for program (S3.6)

$$\mathcal{A} = \{\Phi : \nabla L(\Phi)_{S_0} = 0, \Lambda \succ 0\}.$$

  contains an element $\tilde{\Phi} \in \mathcal{A}$ that satisfies $\|\tilde{\Phi} - \Phi_0\|_\infty \leq r_n$.

- *Step 3:* Prove that $\tilde{\Phi}$ is a local minimizer of the objective function $L(\Phi)$ by showing that $L(\Phi) \geq L(\tilde{\Phi})$ for any Φ in a small neighborhood of $\tilde{\Phi}$.

At *Step 2*, we apply the following lemma. Its proof is deferred to to Section S3.4.2.

Lemma S3.2.1. *Let* $r = \frac{4c_H}{n}\big(\|\nabla \ell(\Phi^0)\|_\infty + pen'_{SS}(\delta_0)\big)$. *If* $\theta^0_{\min} \geq r + \delta_0$ *for some* $\delta_0 > 0$*, and* $r \leq \min\left\{\frac{1}{3c_{\Sigma^0}d}, \frac{1}{3708d^2c^2_{\Gamma^0}c^4_{\Sigma^0}\rho_2}, \frac{c_{\Theta^0}}{2d}\right\}$, *then there exists* $\tilde{\Phi} \in \mathcal{A}$ *such that* $\|\tilde{\Phi} - \Phi^0\|_\infty \leq r$.

Let $\tilde{\Phi}$ denote the local minimizer constructed at *Step 2*. By Lemma S3.2.1, we have $\|\tilde{\Delta}\|_\infty \leq r \leq r_n$ where $\tilde{\Delta} = \tilde{\Phi} - \Phi^0$. Note that Φ^0 satisfies the constraints in program (S3.6)). Therefore $\tilde{\Phi}$ is in the HPD region since $L(\tilde{\Phi}) \leq L(\Phi^0)$.

At *Step 3*, we will show that $\tilde{\Phi}$ is indeed a local minimizer of our objective function $L(\Phi)$ in the unconstrained HPD region. Define

$$\begin{aligned} D(\Delta) &= L(\tilde{\Phi} + \Delta) - L(\tilde{\Phi}) \\ &= \ell(\tilde{\Phi} + \Delta) - \ell(\tilde{\Phi}) + \mathrm{Pen}(\tilde{\Phi} + \Delta) - \mathrm{Pen}(\tilde{\Phi}). \end{aligned}$$

It suffices to show that $D(\Delta) \geq 0$ for any Δ satisfying $\|\Delta\|_\infty \leq \epsilon$ where ϵ is a small positive number we choose, and $\tilde{\Phi} + \Delta$ is in the HPD region.

Since both $\tilde{\Phi}$ and $\tilde{\Phi} + \Delta$ are in the HPD region where $\ell(\Phi)$ is strongly convex at Φ^0. When $r \geq \|\tilde{\Delta}\|_\infty$ and $\epsilon \geq \|\Delta\|_\infty$ are chosen to be small enough, we have

$$\begin{aligned}\ell(\tilde{\Phi} + \Delta) - \ell(\tilde{\Phi}) &\geq \langle \nabla \ell(\tilde{\Phi}), \Delta \rangle + n\beta_0 \|\Delta\|_F^2 \\ &\geq \sum_{i,j \in S_0} \nabla \ell(\tilde{\Phi})_{ij} \Delta_{ij} + \sum_{i,j \in S_0^c} \nabla \ell(\tilde{\Phi})_{ij} \Delta_{ij} + n\beta_0 \|\tilde{\Delta}\|_F^2.\end{aligned} \tag{S3.7}$$

Next we bound $\mathrm{Pen}(\tilde{\Phi} + \Delta) - \mathrm{Pen}(\tilde{\Phi})$. Since $\mathrm{pen}_{\mathrm{SS}}''(\theta) \geq -\mu/2$ for any $|\theta| \geq \delta_0$. Therefore for any $\theta_1, \theta_2 > \delta_0$, we have

$$\mathrm{pen}_{\mathrm{SS}}(\theta_1) - \mathrm{pen}_{\mathrm{SS}}(\theta_2) \geq \mathrm{pen}_{\mathrm{SS}}'(\theta)(\theta_1 - \theta_2) - \frac{\mu}{2}(\theta_1 - \theta_2)^2. \tag{S3.8}$$

since $\min_{(i,j) \in S_0} |\tilde{\Phi}_{ij}| > \delta$ and $|\Delta_{ij}|$ can be made very small, applying (S3.8), we have

$$\begin{aligned}\sum_{(i,j) \in S_0} \mathrm{pen}_{\mathrm{SS}}(\tilde{\Phi}_{ij} + \Delta_{ij}) - \mathrm{pen}_{\mathrm{SS}}(\tilde{\Phi}_{ij}) &\geq \sum_{i,j \in S_0} \mathrm{pen}_{\mathrm{SS}}'(\tilde{\Phi}_{ij}) \Delta_{ij} - \frac{\mu}{2} \|\Delta_{S_0}\|_F^2 \\ &\geq \sum_{i,j \in S_0} \mathrm{pen}_{\mathrm{SS}}'(\tilde{\Phi}_{ij}) \Delta_{ij} - \frac{\mu}{2} \|\Delta\|_F^2.\end{aligned} \tag{S3.9}$$

Since $\tilde{\Phi}_{S_0^c} = 0$ and $\max_{(i,j)\in S_0^c} |\tilde{\Delta}_{ij}| \le \epsilon$,

$$\begin{aligned}\sum_{(i,j)\in S_0^c} \mathrm{pen}_{\mathrm{SS}}(\tilde{\Phi}_{ij} + \Delta_{ij}) - \mathrm{pen}_{\mathrm{SS}}(\tilde{\Phi}_{ij}) &= \sum_{(i,j)\in S_0^c} \mathrm{pen}_{\mathrm{SS}}(\Delta_{ij}) \\ &\ge \sum_{(i,j)\in S_0^c} \mathrm{pen}'_{\mathrm{SS}}(\epsilon)|\Delta_{ij}|\end{aligned} \tag{S3.10}$$

due to the monotonicity of $\mathrm{pen}'(\cdot)$.

Combine (S3.7), (S3.9), and (S3.10) to express $D(\Delta)$ as follows:

$$\begin{aligned} D(\Delta) \;\ge\; & \sum_{i,j\in S_0} \big(\nabla \ell(\tilde{\Phi})_{ij} + \mathrm{pen}'_{\mathrm{SS}}(\tilde{\Phi}_{ij})\big)\cdot \Delta_{ij} + \frac{2n\beta_0 - \mu}{2}\|\Delta\|_F^2 \qquad \text{(I)} \\ & + \sum_{i,j\in S_0^c} \big(\mathrm{pen}'_{\mathrm{SS}}(\epsilon) - |\nabla \ell(\tilde{\Phi})_{ij}|\big)\cdot|\Delta_{ij}| \qquad \text{(II)} \end{aligned}$$

For (I), the first term is equal to zero due to the zero subgradient condition at $\tilde{\Phi}$, which is a (local) minimal of the optimization problem considered at *Step 1.* The second term is positive provided $n\beta_0 > \mu/2$.

Next we aim to show that (II) is positive. Since ϵ can be chosen to be arbitrarily close to 0, it suffices to show $\|\nabla \ell(\tilde{\Phi})_{S_0^c}\|_\infty < \mathrm{pen}'_{\mathrm{SS}}(0^+)$. Apply the first order Taylor expansion of $\nabla \ell(\tilde{\Phi}) = \nabla \ell(\Phi^0 + \tilde{\Delta})$ at Φ^0, namely,

$$\nabla \ell(\tilde{\Phi}) = \nabla \ell(\Phi^0) + H\tilde{\Delta} + R(\tilde{\Delta}), \tag{S3.11}$$

where $R(\tilde{\Delta})$ denotes the residual, on elements from S_0^c. Then we have $\nabla \ell(\tilde{\Phi})_{S_0^c} = \nabla \ell(\Phi^0)_{S_0^c} + H_{S_0^c S_0}\tilde{\Delta}_{S_0} + R(\tilde{\Delta})_{S_0^c}$. Therefore

$$\begin{aligned} \|\nabla \ell(\tilde{\Phi})_{S_0^c}\|_\infty &\le \|\nabla \ell(\Phi^0)\|_\infty + c_H \frac{n}{2}\|\tilde{\Delta}_{S_0}\|_\infty + \|R(\tilde{\Delta})\|_\infty \\ &\le \frac{nr}{2}\Big(\frac{1}{c_H} + c_H\Big) < \mathrm{pen}'_{\mathrm{SS}}(0^+), \end{aligned}$$

since $|||H_{S_0^c S_0}|||_\infty \leq |||H|||_\infty = \frac{n}{2}c_H$, and by the proof of Lemma S3.2.1, $\|\nabla\ell(\Phi^0)\|_\infty + \|R(\tilde{\Delta})\|_\infty \leq \frac{n}{2}r/c_H$ and $\|\tilde{\Delta}_{S_0}\|_\infty = \|\tilde{\Delta}\|_\infty \leq r$. $\square$

S3.3 Proof of Theorem 3

Proof. Note that with the fractional likelihood, the penalty term in the objective function $L_\kappa(\Phi)$ is the original penalty function $\mathrm{Pen}(\Phi)$ scaled by κ. It is easy to check that $\kappa\mathrm{pen}'_{\mathrm{SS}}(\theta)$ and $\kappa\mathrm{pen}''_{\mathrm{SS}}(\theta)$ satisfy the conditions in Theorem 1 and Theorem 2. Further we have

$$\kappa\mathrm{pen}''_{\mathrm{SS}}(\theta) \leq \frac{\kappa}{4}\left(\frac{1}{v_0} - \frac{1}{v_1}\right)^2 \sim O(n).$$

As implied in our proof for Theorem 1, the negative log-likelihood function $-\ell(\Phi)$ is strongly convex in HPD with its second derivative lower bounded by $n\beta_0$. Therefore with a proper choice of the constants, we have $\partial^2 L_\kappa(\Phi) > 0$, therefore $L_\kappa(\Phi)$ is convex in HPD with a unique stationary point.

$\square$

S3.4 Proofs of technical lemmas

S3.4.1 Proof of Lemma 3

Proof. Since $L(\Phi^0) \geq L(\Phi)$, we have

$$\begin{aligned}\mathrm{Pen}(\Phi^0) - \mathrm{Pen}(\Phi) \geq \ell(\Phi) - \ell(\Phi^0) &\geq \langle \nabla \ell(\Phi^0), \Phi - \Phi^0 \rangle \qquad \text{(S3.12)} \\ &\geq -\gamma_n(\|\Delta_{S_0^c}\|_1 + \|\Delta_{S_0}\|_1)\end{aligned}$$

due to the convexity of $\ell(\Phi)$, where $\gamma_n := \|\ell(\Phi^0)\|_\infty$.

$$\begin{aligned}\mathrm{Pen}(\Phi^0) - \mathrm{Pen}(\Phi) &= -\sum_{j \in S_0^c} \mathrm{pen}_{\mathrm{SS}}(\Phi_j) + \sum_{j \in S_0} \Big(\mathrm{pen}_{\mathrm{SS}}(\Phi_j^0) - \mathrm{pen}_{\mathrm{SS}}(\Phi_j^0) \Big) \\ &= -\sum_{j \in S_0^c} \mathrm{pen}_{\mathrm{SS}}(\Delta_j) + \sum_{j \in S_0} \Big(\mathrm{pen}_{\mathrm{SS}}(\Phi_j^0) - \mathrm{pen}_{\mathrm{SS}}(\Phi_j^0 + \Delta_j) \Big) \\ &\leq -\frac{1}{v_1} \|\Delta_{S_0^c}\|_1 + \frac{1}{v_0} \|\Delta_{S_0}\|_1\end{aligned}$$

where the last inequality is due to the following properties of the spike-and-slab Lasso penalty: $|\mathrm{pen}_{\mathrm{SS}}(t)| \geq |t|/v_1$ and $|\mathrm{pen}_{\mathrm{SS}}(t_1) - \mathrm{pen}(t_2)| \leq |t_1 - t_2|/v_0$.

Re-organizing (S3.12), we have $\|\Delta_{S_0^c}\|_1 \leq \alpha \|\Delta_{S_0}\|_1$ with $\alpha = (v_0 + 2v_1)/v_0 \geq (1/v_0 + \gamma_n)/(1/v_1 - \gamma_n)$, provided that $1/v_1 \geq 2\gamma_n$.

$\square$

S3.4.2 Proof of Lemma S3.2.1

Proof. Let $\tilde{\Delta} = \tilde{\Phi} - \Phi^0$. It suffices to show that $\|\tilde{\Delta}_{S_0}\|_\infty \leq r$. Our approach is similar to that of **?** and **?**.

Note that any $\tilde{\Phi} \in \mathcal{A}$ must satisfy the zero-subgradient condition:

$$\nabla \ell(\tilde{\Phi})_{S_0} + Z(\tilde{\Phi})_{S_0} = 0, \tag{S3.13}$$

where $Z(\tilde{\Phi})$ denotes the sub-gradient vector of the penalty term evaluated at $\tilde{\Phi}$ with $Z(\tilde{\Phi})_{ij} = \mathrm{pen}'_{\mathrm{SS}}(|\tilde{\Phi}_{ij}|) \cdot \mathrm{sign}(\tilde{\Phi}_{ij})$, where $\mathrm{sign}(x)$, the (sub)-gradient of $|x|$, equals 1 if $x > 0$, -1 if $x < 0$, and any number from $[-1, 1]$ if $x = 0$.

Equation (S3.13) can be expressed as

$$\nabla \ell(\Phi^0)_{S_0} + H_{S_0 S_0} \cdot \mathrm{vec}(\tilde{\Delta}_{S_0}) + R(\tilde{\Delta})_{S_0} + Z(\tilde{\Phi})_{S_0} = 0 \tag{S3.14}$$

due to (S3.11). The negative log-likelihood function is strongly convex on the constrained set, so $H_{S_0 S_0}$ is invertible and

$$\mathrm{vec}(\tilde{\Delta}_{S_0}) = -H_{S_0 S_0}^{-1} \Big(\nabla \ell(\Phi^0)_{S_0} + R(\tilde{\Delta})_{S_0} + Z(\tilde{\Phi})_{S_0} \Big).$$

Next define a mapping F from $\mathbb{R}^{|S_0|} \to \mathbb{R}^{|S_0|}$:

$$F(\mathrm{vec}(\tilde{\Delta}_{S_0})) := -H_{S_0 S_0}^{-1} \Big(\nabla \ell(\Phi^0)_{S_0} + R(\tilde{\Delta})_{S_0} + Z(\tilde{\Phi})_{S_0} \Big). \tag{S3.15}$$

By construction, $F(\mathrm{vec}(\tilde{\Delta}_{S_0})) = \mathrm{vec}(\tilde{\Delta}_{S_0})$ iff (S3.14) holds. Since F is continuous, if we could show that $\|F(\mathrm{vec}(\tilde{\Delta}_{S_0}))\|_\infty \leq r$ for any $\|\tilde{\Delta}_{S_0}\|_\infty \leq r$, then by Brouwer's fixed point theorem, there exists a solution to program (S3.6), $\tilde{\Phi} \in \mathcal{A}$, and $\|\tilde{\Phi} - \Phi^0\|_\infty = \|\tilde{\Delta}_{S_0}\|_\infty \leq r$.

Bound (S3.15) as follows

$$\|F(\text{vec}(\tilde{\Delta}_{S_0}))\|_\infty \leq c_H\Big(\frac{2}{n}\|\nabla\ell(\Phi^0)_{S_0}\|_\infty + \frac{2}{n}\|R(\tilde{\Delta})_{S_0}\|_\infty + \frac{2}{n}\|Z(\tilde{\Phi})_{S_0}\|_\infty\Big), \tag{S3.16}$$

where the $2/n$ factor comes from $c_H = \frac{n}{2}\|\!|H^{-1}_{S_0S_0}|\!\|_\infty$. We have

$$c_H\frac{2}{n}\|R(\Delta)\|_\infty \leq 1854 d^2 c^2_{\Gamma^0} c^4_{\Sigma^0}\rho_2\|\Delta\|^2_\infty \leq r/2$$

by our condition and Lemma 1. Since $\theta^0_{\min} \geq r + \delta_0$, when $\|\tilde{\Delta}\|_\infty \leq r$, we have $\min_{(i,j)\in S_0}|\tilde{\Delta}_{ij}| \geq \delta_0$. Therefore we can upper bound $\|Z(\tilde{\Phi})_{S_0}\|_\infty$ by $\text{pen}'_{SS}(\delta_0)$. It is easy to check that the right hand side of (S3.16) is bounded by $r/2 + \frac{2c_H}{n}\big(\|\nabla\ell(\Phi^0)\|_\infty + \text{pen}'_{SS}(\delta_0)\big) = r$.

Further, Because $dr < \frac{1}{3c_{\Sigma^0}} < \lambda_{\min}(\Lambda^0)$, $\lambda_{\min}(\tilde{\Lambda}) > \lambda_{\min}(\Lambda^0) - \lambda_{\max}(\tilde{\Lambda} - \Lambda^0) > dr - dr > 0$, $\lambda_{\min}(\tilde{\Lambda}) > 0$. So $\tilde{\Phi} = (\tilde{\Lambda}, \tilde{\Theta})$ is inside $\mathcal{A}$ by assumption, i.e., $\mathcal{A}$ is not empty. $\square$

S4 The EM algorithm

In this section, we work through the details in deriving the EM algorithm for implementing our method. Let $\Phi = (\Lambda, \Theta)$, and R^Λ and R^Θ be binary indicator matrices with their (i,j)th entries being r^Λ_{ij} and r^Θ_{ij}, respectively.

The likelihood function of the Gaussian conditional random field model takes the following form:

$$
\begin{aligned}
p(\mathbf{Y} \mid \Phi, \mathbf{X}) &= \prod_{i=1}^{n} p(Y_i \mid X_i, \Phi) \\
&= \left(\frac{1}{\sqrt{(2\pi)^p |\Lambda^{-1}|^{\frac{1}{2}}}} \right)^n \exp(-\frac{n}{2}\mathrm{tr}(S_{yy}\Lambda + 2S_{xy}\Theta + \Lambda^{-1}\Theta^T S_{xx}\Theta)).
\end{aligned}
\tag{S4.17}
$$

According to our Bayesian model specification, the full posterior distribution $\pi(\Phi, R^{\Phi}|\mathbf{Y}, \mathbf{X})$ is proportional to

$$
p(\mathbf{Y} \mid \Phi, \mathbf{X}) \cdot \Big[\prod_{i,j} \pi(\Phi_{ij}|r_{ij}^{\Phi})\pi(r_{ij}^{\Phi}|\eta) \Big] \cdot \mathbf{1}(\Lambda \succ 0, \|\Lambda\|_2 \leq R).
$$

We propose an efficient EM algorithm by treating R^{Φ} as latent. From the proposed EM algorithm, we will find an optimizer of the loss function

$$
\operatorname*{arg\,min}_{\Theta, \Lambda \succ 0, \|\Lambda\|_2 \leq R} L(\Phi)
$$

along with estimates of the posterior inclusion probabilities of r_{ij}^{Φ}, which will be denoted as p_{ij}^{Φ}. The posterior probabilities p_{ij}^{Φ} can be used to determine the sparse pattern of Φ_{ij}. Due to our theoretical results on estimation accuracy, thresholding the posterior probabilities p_{ij}^{Φ} will lead to consistent structure recovery when the minimal signal strength is strong enough. The proposed EM algorithm for the GCRF model is motivated by the EMVS algorithm for linear regression (?) and its variants for factor model and Gaussian graphical model settings (??).

The E-step

Compute the posterior of R^{Φ} given the estimate of Φ from the previous iteration. It can be shown that $r^{\Phi}_{ij}|\Phi^{(t)}_{ij}$ follows (p^{Φ}_{ij}) where

$$\log \frac{p^{\Phi}_{ij}}{1-p^{\Phi}_{ij}} = \Big(\log \frac{v_0}{v_1} + \log \frac{\eta}{1-\eta} - \frac{|\Phi^{(t)}_{ij}|}{v_1} + \frac{|\Phi^{(t)}_{ij}|}{v_0} \Big). \tag{S4.18}$$

Next, compute the expectation of the log posterior with respect to R^{Φ}, which gives rise to the Q function to be maximized:

$$Q(\Phi|\Phi^{(t)}) = \ell(\Phi) - \sum_{i,j} \lambda(\Phi^{(t)}_{ij}) \cdot |\Phi_{ij}|, \quad \lambda(\Phi^{(t)}_{ij}) = \frac{p^{\Phi}_{ij}}{v_1} + \frac{1-p^{\Phi}_{ij}}{v_0}, \tag{S4.19}$$

where $\ell(\Phi)$ is the log-likelihood function defined in (S2.1).

The M-step

Optimizing (S4.19) with respect to $\Phi = (\Theta, \Lambda)$ is equivalent to optimizing GCRF with a *weighted* Lasso penalty. We implemented a second order algorithm, motivated by the algorithm from ? for the usual unweighted Lasso penalty.

At the beginning of each iteration in the M-step, we first form a second order approximation of $\ell(\Theta+\Delta_\Theta, \Lambda+\Delta_\Lambda)$ at (Θ, Λ),

$$
\begin{aligned}
g(\Delta_\Theta, \Delta_\Lambda) = \ell(\Theta,\Lambda) &+ \mathrm{vec}(\nabla \ell(\Theta,\Lambda))^T \mathrm{vec}(\Delta_\Phi) \\
&+ \frac{1}{2}\mathrm{vec}(\Delta_\Phi)\mathrm{vec}(\nabla^2 \ell(\Theta,\Lambda))^T \mathrm{vec}(\Delta_\Phi),
\end{aligned} \tag{S4.20}
$$

where $\Delta_\Phi = \begin{bmatrix} \Delta_\Lambda \\ \\ \Delta_\Theta \end{bmatrix}$, and the gradient $\nabla_\Theta \ell(\Theta, \Lambda)$ and Hessian matrix $\nabla^2 \ell(\Theta, \Lambda)$ of the log-likelihood function $\ell(\Theta, \Lambda)$ are defined in Appendix S2. At each iteration, we estimate the direction Δ_Φ based on the second order approximation (S4.20):

$$
\begin{aligned}
\arg\max_{\Delta_\Phi} \Big\{ g(\Delta_\Theta, \Delta_\Lambda) &- \sum_{\substack{i\in(1,2,..,p)\\ j\in(1,2,..,p)\\ i<j}} \Big[(1-p_{ij}^{\Lambda})\frac{|\Lambda_{ij}+\Delta_{\Lambda ij}|}{v_0} + p_{ij}^{\Lambda}\frac{|\Lambda_{ij}+\Delta_{\Lambda ij}|}{v_1} \Big] \\
&- \sum_{\substack{i\in(1,2,..,q)\\ j\in(1,2,..,p)}} \Big[(1-p_{ij}^{\Theta})\frac{|\Theta_{ij}+\Delta_{\Theta ij}|}{v_0} + p_{ij}^{\Theta}\frac{|\Theta_{ij}+\Delta_{\Theta ij}|}{v_1} \Big] \Big\}.
\end{aligned} \tag{S4.21}
$$

We use cyclic coordinate descent approach to estimate Δ_Φ. Once we solve the optimization problem for coordinate Φ_{ij}, which results in the Newton direction D_{ij}, we update Δ_Φ by $\Delta_\Phi \leftarrow \Delta_\Phi + D_{ij}$. We iterate over all the coordinates of Φ to get the full updating direction Δ_Φ of an M-step.

The cyclic coordinate descent approach for elements of Δ_Φ can be divided into three sub-problems:

1. Update for the entries in Θ;

2. Update for the off-diagonal entries in Λ;

3. Update for the diagonal entries in Λ.

Each of the subproblems can be written as a simple Lasso problem in the following form:

$$\arg\min_{d}\left(\frac{1}{2}ad^2+bd+\lambda|c+d|\right), \tag{S4.22}$$

with appropriate definitions for a, b, c, and λ in each case. The above objective function has a closed form solution given by

$$d=-c+S_{\lambda/a}\Big(c-\frac{b}{a}\Big), \tag{S4.23}$$

where $S_\lambda(x)=\operatorname{sign}(x)\max(|x|-\lambda,0)$. We shall now provide explicit expressions for (a,b,c,λ) in each of the three cases above.

Update entries in Θ: We can decouple the optimization problem (S4.21) of Θ_{ij} as follows:

$$\begin{aligned}
&\arg\max_{u}\frac{n}{2}\Big[-\operatorname{tr}\Big((2S_{xy}+2S_{xx}\Theta\Lambda^{-1}+2S_{xx}\Delta_\Theta\Lambda^{-1})ue_ie_j^T\\
&\qquad+2(\Delta_\Theta+ue_ie_j^T)\Lambda^{-1}\Delta_\Lambda\Lambda^{-1}\Theta^TS_{xx}-\Lambda^{-1}(\Delta_\Theta+ue_ie_j^T)S_{xx}(\Delta_\Theta+ue_ie_j^T)\Big)\Big]\\
&\qquad-\Big[\frac{1-p_{ij}^{\Theta}}{v_0}+\frac{p_{ij}^{\Theta}}{v_1}\Big]|\Theta_{ij}+(\Delta_\Theta)_{ij}+u|\\
=&\arg\max_{u}\frac{n}{2}\Big[-(2S_{xy}+2S_{xx}\Theta\Lambda^{-1}+2S_{xx}\Delta_\Theta\Lambda^{-1}-2\Lambda^{-1}\Delta_\Lambda\Lambda^{-1}\Theta^TS_{xx})_{ij}u-u^2(\Lambda^{-1})_{jj}(S_{xx})_{ii}\Big]\\
&\qquad-\Big[\frac{1-p_{ij}^{\Theta}}{v_0}+\frac{p_{ij}^{\Theta}}{v_1}\Big]|\Theta_{ij}+(\Delta_\Theta)_{ij}+u|.
\end{aligned}$$

Define a, b, c, λ as follows:

$$\begin{cases} a = n(\Lambda^{-1})_{jj}(S_{xx})_{ii} \\ b = n(S_{xy} + S_{xx}\Theta\Lambda^{-1} + S_{xx}\Delta_{\Theta}\Lambda^{-1} - 2\Lambda^{-1}\Delta_{\Lambda}\Lambda^{-1}\Theta^{T}S_{xx})_{ij} \\ c = \Theta_{ij} + (\Delta_{\Theta})_{ij} \\ \lambda = \frac{1-p_{ij}^{\Theta}}{v_0} + \frac{p_{ij}^{\Theta}}{v_1}, \end{cases} \tag{S4.24}$$

we solve the updating direction w.r.t Θ_{ij} through (S4.23).

Update the off-diagonal entries in Λ: We decompose the loss function in the optimization problem and extract the function only about Λ_{ij}. Denote $A = \Lambda^{-1}\Theta^{T}S_{xx}\Theta\Lambda^{-1}$ and $B = \Lambda^{-1} + 2A$. The optimization problem about off-diagonal entry Λ_{ij} is:

$$
\begin{aligned}
&\arg\max_u \frac{n}{2}\Big[\mathrm{tr}\Big((-S_{yy}+\Lambda^{-1}+A)u(e_ie_j^T+e_je_i^T) \\
&\qquad + 2(\Delta_\Lambda + u(e_ie_j^T+e_je_i^T))\Lambda^{-1}\Delta_\Theta\Lambda^{-1}\Theta^T S_{xx} \\
&\qquad - \Lambda^{-1}(\Delta_\Lambda + u(e_ie_j^T+e_je_i^T)B(\Delta_\Lambda + u(e_ie_j^T+e_je_i^T)\Big)\Big] \\
&\qquad - \Big[\frac{1-p_{ij}^\Lambda}{v_0}+\frac{p_{ij}^\Lambda}{v_1}\Big]|\Lambda_{ij}+(\Delta_\Lambda)_{ij}+u| \\
=&\arg\max_u \frac{n}{2}\Big(-\Big[(\Lambda_{ij}^{-1})^2+\Lambda_{ii}^{-1}\Lambda_{jj}^{-1}+\Lambda_{ii}^{-1}A_{jj}+2\Lambda_{ij}^{-1}A_{ij}+\Lambda_{jj}^{-1}A_{ii}\Big]u^2 \\
&\qquad + 2\Big[(-S_{yy}+\Lambda^{-1}+A)_{ij}+(\Lambda^{-1}\Delta_\Theta\Lambda^{-1}\Theta^T S_{xx})_{ij}+(\Lambda^{-1}\Delta_\Theta\Lambda^{-1}\Theta^T S_{xx})_{ji} \\
&\qquad - (\Lambda^{-1}\Delta_\Lambda\Lambda^{-1})_{ij}-(\Lambda^{-1}\Delta_\Lambda A)_{ij}-(\Lambda^{-1}\Delta_\Lambda A)_{ji}\Big]u\Big) \\
&\qquad - \Big[\frac{1-p_{ij}^\Lambda}{v_0}+\frac{p_{ij}^\Lambda}{v_1}\Big]|\Lambda_{ij}+(\Delta_\Lambda)_{ij}+u|.
\end{aligned}
\tag{S4.25}
$$

To update the off-diagonal entries Λ_{ij}, we define a, b, c, λ in equation (S4.22) as

$$
\begin{cases}
a = n\Big[(\Lambda_{ij}^{-1})^2+\Lambda_{ii}^{-1}\Lambda_{jj}^{-1}+\Lambda_{ii}^{-1}A_{jj}+\Lambda_{ij}^{-1}A_{ij}+\Lambda_{jj}^{-1}A_{ii}\Big] \\
b = -n\Big[(-S_{yy}+\Lambda^{-1}+A)_{ij}+(\Lambda^{-1}\Delta_\Theta\Lambda^{-1}\Theta^T S_{xx})_{ij}+(\Lambda^{-1}\Delta_\Theta\Lambda^{-1}\Theta^T S_{xx})_{ji} \\
\qquad -(\Lambda^{-1}\Delta_\Lambda\Lambda^{-1})_{ij}-(\Lambda^{-1}\Delta_\Lambda A)_{ij}-(\Lambda^{-1}\Delta_\Lambda A)_{ji}\Big] \\
c = \Lambda_{ij}+(\Delta_\Lambda)_{ij} \\
\lambda = \frac{1-p_{ij}^\Lambda}{v_0}+\frac{p_{ij}^\Lambda}{v_1},
\end{cases}
\tag{S4.26}
$$

and solve the updating direction w.r.t the coordinate of Λ_{ij} with (S4.23).

Update the diagonal entries in Θ: For the diagonal entries Λ_{ii}, the decoupled optimization problem has the following form:

$$\arg\max_u \frac{n}{2}\Big(-\frac{1}{2}B_{ii}u^2 + \Big[-S_{yy} + \Lambda^{-1} + A + 2(\Lambda^{-1}\Delta_\Theta\Lambda^{-1}\Theta^T S_{xx})_{ii} - (\Lambda^{-1}\Delta_\Lambda\Lambda^{-1})_{ii} - 2(\Lambda^{-1}\Delta_\Lambda A)_{ii}\Big]u\Big), \tag{S4.27}$$

So here we define a, b, c, λ as:

$$\begin{cases} a = \frac{n}{2}B_{ii}, \\ b = -\frac{n}{2}\Big[-S_{yy} + \Lambda^{-1} + A + 2(\Lambda^{-1}\Delta_\Theta\Lambda^{-1}\Theta^T S_{xx})_{ii} \\ \qquad -(\Lambda^{-1}\Delta_\Lambda\Lambda^{-1})_{ii} - 2(\Lambda^{-1}\Delta_\Lambda A)_{ii}\Big], \\ c = \Lambda_{ii} + (\Delta_\Lambda)_{ii}, \\ \lambda = 0. \end{cases} \tag{S4.28}$$

After solving Δ_Φ from problem (S4.21), we update the estimate Φ by $\Phi \leftarrow \Phi + \alpha\Delta_\Phi$, where α is the step size determined by checking (1) Armijo's rule, (2) positive definiteness and (3) boundedness of the resultant matrix. The pseudo-code for this algorithm is in Algorithm 1.

A naive implementation of matrix multiplications for $\Delta_\Theta\Lambda^{-1}$ and $\Delta_\Lambda\Lambda^{-1}$ has a cost of $O(\max(p^3, qp^2))$. Since we need to solve $O(\max(p^2, pq))$ sub-problems when solving (S4.21) and each sub-problem requires to evaluate

Algorithm 1 EM algorithm

Input X, Y.

Initialize $\Theta = \mathit{0}, \Lambda = I$.

While (not reached convergence) do:

(E-Step:)

Calculate P^{Θ}, P^{Λ} using (S4.18).

(M-Step:)

While (not reached convergence) do:

Compute the newton direction Δ_{Φ}.

Determine the step size α by backtracking line search.

Update $\Phi \leftarrow \Phi + \alpha\Delta_{\Phi}$.

Output Φ.

$\Delta_{\Theta}\Lambda^{-1}$ and $\Delta_{\Lambda}\Lambda^{-1}$, the total cost of our algorithm, if using the naive implementation, could be as high as $O(\max(p^5, q^2p^3))$. Inspired by the algorithm from **?**, we developed a more efficient implementation of our method by utilizing the facts that *i*) only one entry in either Δ_{Θ} or Δ_{Λ} is updated when solving each sub-problem, and *ii*) only one row of matrix product $\Delta_{\Theta}\Lambda^{-1}$ or $\Delta_{\Lambda}\Lambda^{-1}$ is updated after each sub-problem. Thus, it is efficient to calculate

$\Delta_\Theta \Lambda^{-1}$ and $\Delta_\Lambda \Lambda^{-1}$ at the beginning of an M-step and then only update the row of matrix product that is changed after solving each sub-problem. The computation cost for this updating scheme is $O(\max(p, q))$, thus our algorithm has a computational cost of $O\left(\max(p^3, p^2 q)\right)$, which matches with the second order algorithm from **?**.

S5 Simulation Results

In this subsection, we provide all the simulation results for the settings described in the main paper.

Table 1: Random Graph Model: Performance comparison of different methods. Larger values of MCC indicate better performance while smaller values of Fnorm and Test Error indicate better performance. Best performing method is highlighted in boldface.

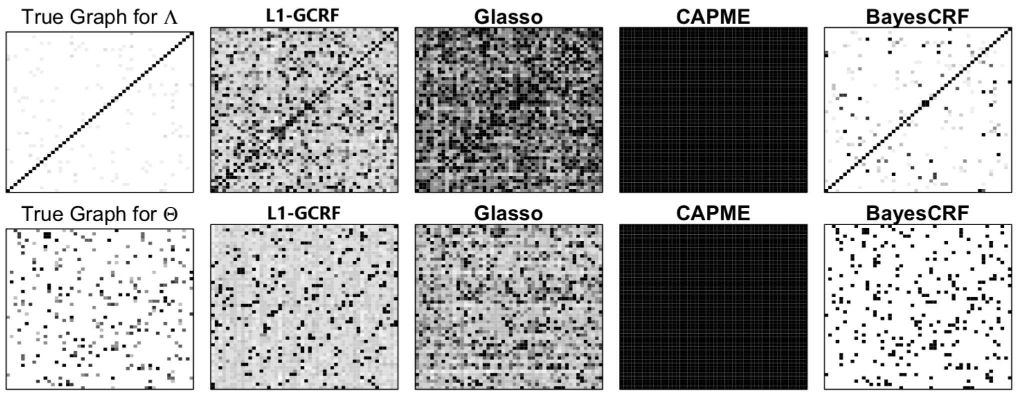
	$n = 100, q = 50, p = 50$			$n = 100, q = 100, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.263(0.039)	10.606(0.735)	2.001(0.296)	0.375(0.013)	17.767(0.061)	4.922(0.181)
CAPME	-0.025(0.001)	46.965(5.653)	2.442(0.125)	-0.020(0.010)	51.674(5.724)	3.934(0.199)
L1-GCRF	0.360(0.0181)	6.901(0.344)	1.446(0.036)	0.481(0.011)	11.709(0.360)	1.652(0.039)
BayesCRF	**0.608(0.010)**	**6.012(0.149)**	**1.390(0.031)**	**0.711(0.006)**	**11.088(0.154)**	**1.560(0.041)**
	$n = 100, q = 200, p = 50$			$n = 100, q = 500, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.337(0.007)	25.472(0.004)	8.180(0.154)	0.180(0.004)	38.747(0.004)	10.366(0.310)
CAPME	-0.015(0.008)	21.532(0.544)	5.433(0.205)	0.000(0.008)	37.889(0.155)	10.086(0.329)
L1-GCRF	0.411(0.008)	22.213(0.338)	**3.142(0.071)**	**0.270(0.012)**	38.963(0.018)	21.706(3.835)
BayesCRF	**0.517(0.036)**	**21.075(0.242)**	3.484(0.601)	0.186(0.008)	**37.127(0.110)**	**7.142(1.341)**

Table 2: Banded Model 1: Performance comparison of different methods. Larger values of MCC indicate better performance while smaller values of Fnorm and Test Error indicate better performance. Best performing method is highlighted in boldface.

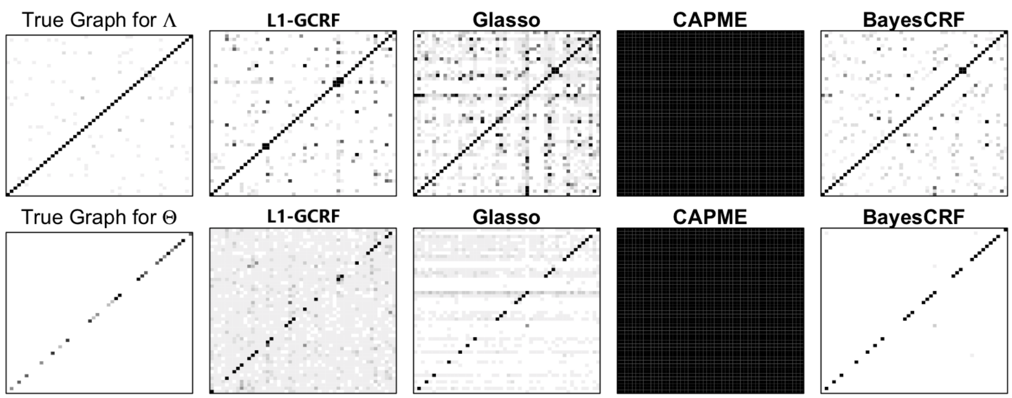
	$n = 100, q = 50, p = 50$			$n = 100, q = 100, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.330(0.022)	4.223(0.040)	1.279(0.032)	0.314(0.015)	5.316(0.035)	1.390(0.035)
CAPME	-0.037(0.001)	30.346(2.709)	1.455(0.046)	-0.036(0.012)	43.642(3.320)	1.696(0.046)
L1-GCRF	0.130(0.020)	3.050(0.110)	**1.250(0.028)**	0.216(0.021)	3.595(0.194)	**1.309(0.031)**
BayesCRF	**0.409(0.026)**	**2.498(0.094)**	1.278(0.032)	**0.452(0.024)**	**2.453(0.077)**	1.335(0.031)
	$n = 100, q = 200, p = 50$			$n = 100, q = 500, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.394(0.012)	9.118(0.015)	2.051(0.053)	0.304(0.046)	12.684(0.162)	2.777(0.187)
CAPME	-0.033(0.010)	63.073(6.914)	2.294(0.069)	0.071(0.004)	13.735(1.546)	2.232(0.060)
L1-GCRF	0.361(0.015)	5.369(0.228)	1.489(0.031)	0.412(0.011)	8.628(0.333)	1.665(0.041)
BayesCRF	**0.606(0.015)**	**3.163(0.110)**	**1.431(0.032)**	**0.674(0.011)**	**6.297(0.143)**	**1.555(0.035)**

Table 3: Banded Model 2: Performance comparison of different methods. Larger values of MCC indicate better performance while smaller values of Fnorm and Test Error indicate better performance. Best performing method is highlighted in boldface.

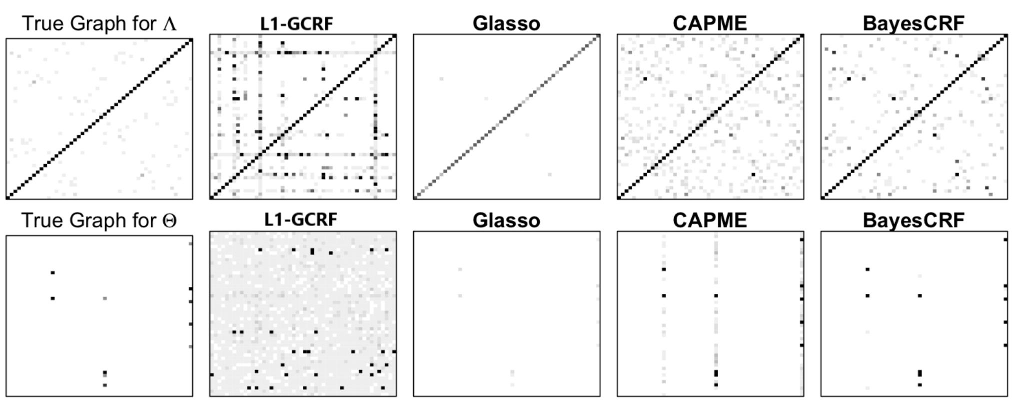
	$n = 100, q = 50, p = 50$			$n = 100, q = 100, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.262(0.017)	3.763(0.047)	1.191(0.031)	0.278(0.015)	5.294(0.031)	1.342(0.030)
CAPME	-0.037(0.000)	27.884(2.113)	1.362(0.044)	-0.035(0.011)	43.030(3.666)	1.658(0.062)
L1-GCRF	0.131(0.023)	3.827(0.136)	**1.215(0.026)**	0.164(0.023)	4.435(0.122)	**1.260(0.027)**
BayesCRF	**0.322(0.026)**	**2.725(0.092)**	1.238(0.031)	**0.392(0.021)**	**2.873(0.106)**	1.316(0.030)
	$n = 100, q = 200, p = 50$			$n = 100, q = 500, p = 50$		
	MCC	Fnorm	Test Error	MCC	Fnorm	Test Error
GLasso	0.326(0.022)	8.489(0.182)	1.775(0.067)	0.255(0.005)	12.543(0.011)	2.577(0.072)
CAPME	-0.034(0.010)	67.937(6.744)	2.066(0.086)	0.109(0.005)	12.534(0.905)	2.166(0.075)
L1-GCRF	0.263(0.017)	6.468(0.119)	**1.379(0.036)**	0.383(0.012)	10.182(0.173)	1.666(0.042)
BayesCRF	**0.476(0.016)**	**3.566(0.097)**	1.386(0.030)	**0.634(0.012)**	**6.372(0.142)**	**1.550(0.038)**



(a) Estimates for Random Graph.



(b) Estimates for Banded Model 1.



(c) Estimates for Banded Model 2.

Figure 1: Visualization of the averages of the estimated graphs when $p = q = 50$. White represents the noise and black represents the selected signal.

References

Gan, L., Narisetty, N. N., and Liang, F. (2019). Bayesian regularization for graphical models with unequal shrinkage. *J. Amer. Statist. Assoc.*, 114(527):1218–1231.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *J. Amer. Statist. Assoc.*, 109(506):828–846.

Ročková, V. and George, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J. Amer. Statist. Assoc.*, 111(516):1608–1622.

Wytock, M. and Kolter, Z. (2013). Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *ICML-13*, pages 1265–1273.

Yuan, X.-T. and Zhang, T. (2014). Partial Gaussian graphical model estimation. *IEEE Transactions on Information Theory*, 60(3):1673–1687.