Statistica Sinica **32** (2022), 1-22
doi:https://doi.org/10.5705/ss.202020.0118

BAYESIAN ESTIMATION OF GAUSSIAN CONDITIONAL RANDOM FIELDS

Lingrui Gan, Naveen N. Narisetty, and Feng Liang

University of Illinois at Urbana-Champaign

Abstract: We propose a novel methodology based on a Bayesian Gaussian conditional random field model for elegantly learning the conditional dependence structures among multiple outcomes, and between the outcomes and a set of covariates simultaneously. Our approach is based on a Bayesian hierarchical model using a spike and slab Lasso prior. We investigate the corresponding maximum a posteriori (MAP) estimator that requires dealing with a nonconvex optimization problem. In spite of the nonconvexity, we establish the statistical accuracy for all points in the high posterior region, including the MAP estimator, and propose an efficient EM algorithm for the computation. Using simulation studies and a real application, we demonstrate the competitive performance of our method for the purpose of learning the dependence structure.

Key words and phrases: Bayesian regularization, Gaussian conditional random field, graphical models, spike and slab Lasso prior.

1. Introduction

Graphical models are widely used in applications where the key interest is to identify the conditional dependence structure among a set of variables $Y = (Y^{(1)}, \ldots, Y^{(p)}) \in \mathbb{R}^p$. A special class of graphical models is the Gaussian graphical model (GGM), under which Y follows a multivariate Gaussian distribution with mean zero and precision matrix Θ. Estimating the underlying dependence structure of a GGM is equivalent to estimating Θ, because it is well known that the (i, j)th element of Θ being zero is equivalent to the conditional independence of $Y^{(i)}$ and $Y^{(j)}$, given the other variables. Owing to this connection, sparse precision matrix estimation is an important and well-studied research problem (Meinshausen and Bühlmann (2006); Banerjee, El Ghaoui and d'Aspremont (2008); Friedman, Hastie and Tibshirani (2008); Rothman, Levina and Zhu (2010); Ravikumar et al. (2011); Gan, Narisetty and Liang (2019)).

In many application contexts, a marginal Gaussian graphical model for the outcomes alone is not sufficient, and it is important to consider covariate infor-

Corresponding author: Naveen N. Narisetty, Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA. E-mail: naveen@illinois.edu.

mation. For example, in the analysis of gene expression data, it is of interest to model genetic outcomes, given biomarker information, and in the context of portfolio analysis, it is of interest to model asset prices, given historical pricing information. In such applications, along with understanding the dependence relationship among the many outcome variables Y, it is also important to study the relationship between Y and the covariates $X = (X^{(1)}, \ldots, X^{(q)}) \in \mathbb{R}^q$. Although one can model (Y, X) jointly using a GGM to obtain the conditional relationship between the Y's, and between the X's and the Y's as a partial product of the model, it is redundant to model the dependence structure among the X's, which leads to inefficiency when $q \gg p$. We discuss this issue in Section 2.

To learn the conditional dependence structures between the outcomes, and between the outcomes and the covariates, Gaussian conditional random field (GCRF) model has been recently considered (Sohn and Kim (2012); Yuan and Zhang (2014); Wytock and Kolter (2013)). The GCRF model provides a more suitable and precise description of the desired conditional dependence structure compared to modeling the entire Gaussian graphical model on both X and Y, or modeling only the dependence structure among Y by eliminating the effects of X using a multivariate regression model (Cai et al. (2012); Rothman, Levina and Zhu (2010); Yin and Li (2011); Deshpande, Ročková and George (2017)). Estimation methods based on an ℓ_1-penalization for the GCRF model have been proposed, and their theoretical properties for estimation accuracy have been studied by Wytock and Kolter (2013) and Yuan and Zhang (2014). A GCRF estimation using an ℓ_1-penalization for latent X is studied by Frot, Jostins and McVean (2019). Although an ℓ_1-penalty encourages sparsity, while being convex, it has some well-known limitations, such as the bias it induces for large parameter values (Fan and Li (2001); Lam and Fan (2009); Zhang (2010); Zhang and Zhang (2012); Loh and Wainwright (2017)). Moreover, the theoretical results for the structure recovery of an ℓ_1-penalization-based GCRF require restrictive mutual incoherence conditions (Wytock and Kolter (2013)). In this paper, we provide an alternative framework for estimating the Gaussian conditional random field model using a Bayesian framework with spike and slab Lasso priors (Ročková (2018); Ročková and George (2018)). The maximum a posteriori (MAP) estimator can be viewed as a penalized likelihood estimator with a nonconvex penalty function induced by the spike and slab Lasso prior. This has been found to have good regularization properties in the contexts of linear regression (Ročková (2018); Ročková and George (2018)) and Gaussian graphical models (Gan, Narisetty and Liang (2019)).

We address novel theoretical and computational challenges posed by the

GCRF model under the Bayesian setting. The likelihood corresponding to the GCRF model need not satisfy the restricted strong convexity property (Loh and Wainwright (2017)). Furthermore, the Bayesian penalty function corresponding to the spike and slab Lasso prior need not have a bounded second derivative for all the parameter values. These are new challenges related to studying the properties of our MAP estimator. For example, without such properties, local optima may not be unique, and general results from existing works (Loh and Wainwright (2017)) on support recovery for nonconvex optimization are not applicable. Despite the challenges imposed by both the likelihood and the nonconvexity, we show that all points from the high posterior density (HPD) region, including the MAP estimator, have an optimal convergence rate in the Frobenius norm. In addition, we show that there exists at least one local optimum that converges in the ℓ_∞-norm and achieves support-recovery consistency, without the incoherence condition required by Wytock and Kolter (2013). We also show that the optimal convergence rate in the ℓ_∞-norm holds for all local modes of the fractional posterior, that is, the posterior defined with respect to a fractional likelihood. Our theoretical results (presented in Section 3) are stronger than those on the Gaussian conditional random field models with an ℓ_1-penalty of Yuan and Zhang (2014) and Wytock and Kolter (2013). More generally, our results provide novel contributions to the theoretical properties of nonconvex penalization, in the spirit of Fan and Li (2001), Lam and Fan (2009), Negahban et al. (2009), Zhang (2010), Zhang and Zhang (2012), Loh and Wainwright (2015), and Loh and Wainwright (2017).

We propose an efficient EM algorithm for the computation (described in the Supplementary Material) that has the same computational complexity as the state-of-the-art optimization algorithm for the Gaussian conditional random field with an ℓ_1-penalty (Wytock and Kolter (2013); Yuan and Zhang (2014)). Our empirical studies in Section 4 demonstrate that the proposed Bayesian regularization approach provides competitive performance compared with that of alternative methods, both for estimation and structure recovery.

2. Bayesian Regularization for Gaussian Conditional Random Fields

2.1. Model formulation

Consider a p-dimensional outcome Y and a q-dimensional covariate X. As an analog to the conditional random field for discrete variables proposed by Lafferty, McCallum and Pereira (2001), the Gaussian conditional random field model (Sohn and Kim (2012); Yuan and Zhang (2014); Wytock and Kolter (2013)) as-

sumes the following conditional density function of Y, given X:

$$p(Y \mid X, \Lambda, \Theta) \propto \sqrt{\det(\Lambda)} \exp\left\{-\frac{1}{2} Y^T \Lambda Y - X^T \Theta Y\right\}, \tag{2.1}$$

where Λ is a $p \times p$ positive-definite and symmetric matrix, and $\Theta \in \mathbb{R}^{q \times p}$ is a matrix of dimension $q \times p$. Throughout, we use Φ as a compact notation for parameters Λ and Θ. Given a set of n random samples $(X_i, Y_i)_{i=1}^n$, the corresponding log-likelihood function is given by

$$\ell(\Phi) = \frac{n}{2}\Big(\log\det(\Lambda) - \mathrm{tr}(S_{yy}\Lambda + 2S_{xy}\Theta + \Lambda^{-1}\Theta^T S_{xx}\Theta)\Big), \tag{2.2}$$

where $S_{yy} = (1/n)\sum_{i=1}^n Y_i Y_i^T$, $S_{xy} = (1/n)\sum_{i=1}^n X_i^T Y_i$, $S_{xx} = (1/n)\sum_{i=1}^n X_i X_i^T$, and the constant terms not involving the parameters are omitted. Irrelevant to the marginal distribution of X, the sparsity patterns of Φ determine the conditional dependence relationship between the components of Y and the dependence between X and Y:

$$\begin{aligned} \Theta_{ij} = 0 \quad &\Longleftrightarrow \quad X^{(i)} \perp\!\!\!\perp Y^{(j)} \mid X^{-(i)}, Y^{-(j)}, \\ \Lambda_{ij} = 0 \quad &\Longleftrightarrow \quad Y^{(i)} \perp\!\!\!\perp Y^{(j)} \mid X, Y^{-(i,j)}, \end{aligned}$$

where $\perp\!\!\!\perp$ denotes independence. Moreover, the GCRF model avoids modeling the dependence structure among the X's, which is beneficial both computationally and theoretically when the dimension of X is large. We now discuss two alternative modeling frameworks that produce descriptions of the conditional dependence structure.

2.1.1. Joint Gaussian graphical model on (X, Y)

One common approach used to learn the dependence structure is to model (X, Y) using a joint graphical model. With the additional assumption that X is normally distributed with mean zero, the GCRF model implies that (X, Y) jointly follows a multivariate Gaussian distribution,

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \Omega_{xx} & \Theta \\ \Theta^T & \Lambda \end{bmatrix}^{-1}\right). \tag{2.3}$$

Therefore Θ and Λ can be obtained as a partial outcome from fitting a large Gaussian graphical model jointly on (X, Y) using existing algorithms on high-dimensional Gaussian graphical models such as the graphical Lasso as done by Witten and Tibshirani (2009).

This approach, however, is not optimal if we are only interested in Θ and Λ. When the dimension of X is much larger than the dimension of Y, the computational cost is dominated by estimating the graphical structure of X, which is not of interest. Theoretically, the error from estimating Ω_{xx} may affect the estimation accuracy when estimating Θ and Λ, because the accuracy is affected by the degree of sparsity of the entire graph (Bickel and Levina (2008); Cai, Liu and Luo (2011); Ravikumar et al. (2011); Loh and Wainwright (2015, 2017); Gan, Narisetty and Liang (2019)). To avoid estimating the irrelevant structure among the X variables, one can work with the profile likelihood

$$\tilde{\ell}(\Phi) = \max_{\Omega_{xx}} \tilde{\ell}(\Omega_{xx}, \Theta, \Lambda),$$

where $\tilde{\ell}(\Omega_{xx}, \Theta, \Lambda) = \log \prod_{i=1}^{n} p(X_i, Y_i \mid \Omega_{xx}, \Theta, \Lambda)$ is the log-likelihood of the joint Gaussian distribution (2.3). As shown by Yuan and Zhang (2014), the profile likelihood $\tilde{\ell}(\Phi)$ is exactly equal to the GCRF likelihood defined by (2.2). Although it can be viewed as the profile likelihood of a joint Gaussian graphical model on (X, Y), our GCRF model makes no assumption on the marginal distribution of X, and is even applicable when X is discrete.

2.1.2. Covariate-adjusted graphical model

The other alternative modeling framework that can be used to learn the conditional dependence structure is the multivariate regression framework. The conditional distribution of Y given X from the GCRF model (2.1) can be reparametrized as a multivariate regression model with B as the regression coefficient matrix, and Λ as the error precision matrix, as follows:

$$Y \mid X \sim N(BX, \Lambda^{-1}), \quad B = -\Lambda^{-1}\Theta^T. \tag{2.4}$$

Within this regression framework, referred to as the covariate-adjusted graphical model, several approaches have been proposed to estimate B and Λ under sparsity assumptions (Cai et al. (2012); Rothman, Levina and Zhu (2010); Yin and Li (2011); Deshpande, Ročková and George (2017)).

Although Λ indeed reveals the conditional dependence structure among the elements of Y, the sparsity pattern of B is different to that of Θ. The regression coefficients B_{ij} indicate how the conditional mean $\mathbb{E}(Y^{(i)}|X)$ depends on the X variables, without conditioning on the other Y variables. In contrast, Θ_{ij} reflects the conditional dependence between $Y^{(i)}$ and $X^{(j)}$, given all other X and Y variables. Apart from the differences in the sparsity structures, another major difference between the two parameterizations is that the log-likelihood function

of the GCRF model parameterized by (Θ, Λ) is convex, whereas that from the multivariate regression, parameterized by (B, Λ), is nonconvex (Yuan and Zhang (2014)).

2.2. Proposed Bayesian regularization formulation

Our goal is to estimate the parameters Θ and Λ for the GCRF model (2.1) under the assumption of sparsity. Although ℓ_1-regularization is a natural choice for the GCRF model as considered by Yuan and Zhang (2014) and Wytock and Kolter (2013), this approach induces bias on the parameters with large values. Furthermore, it requires strong mutual incoherence assumptions for consistent graph structure recovery. This motivates us to consider an alternative formulation from the Bayesian regularization framework owing to its promising performance in recent works (Ročková and George (2016, 2018); Gan, Narisetty and Liang (2019)).

We consider the spike and slab Lasso prior, which takes the form of a mixture of two Laplace distributions:

$$\pi_{\text{SS}}(\theta) = \eta \cdot \text{LP}(\theta; v_1) + (1 - \eta) \cdot \text{LP}(\theta; v_0), \tag{2.5}$$

where $\text{LP}(\theta; v) = 1/(2v)e^{-|\theta|/v}$ denotes the density function of a Laplace distribution with scale parameter v, the two scale parameters satisfy $v_1 > v_0 > 0$, and η is the mixing weight. Spike and slab priors with Gaussian components have long been used for Bayesian variable selection (George and McCulloch (1993); Ishwaran and Rao (2005); Narisetty and He (2014)). More recently, the spike and slab Lasso prior has been shown to yield desirable shrinkage properties for sparse estimation (Ročková and George (2014); Ročková (2018); Ročková and George (2018); Gan, Narisetty and Liang (2019)).

The following alternative representation of the spike and slab Lasso prior (2.5) may help to explain the motivation behind such a mixture representation:

$$\pi(\theta|\gamma) = \text{LP}(\theta; v_1)^{\gamma} \cdot \text{LP}(\theta; v_0)^{(1-\gamma)}, \quad \gamma \sim \mathsf{Bern}(\eta),$$

where the binary variable γ can be interpreted as the indicator for θ being a signal or noise. When $\gamma = 1$, the unknown parameter θ is expected to represent a signal taking a relatively large value, and is modeled by a Laplace distribution with a larger scale parameter v_1 (i.e., the "slab" component); when $\gamma = 0$, the unknown parameter θ is expected to represent noise, taking a value close to zero, and is modeled by a Laplace distribution with a small scale parameter v_0 (i.e., the "spike" component).

We place the spike and slab Lasso prior on all entries of Θ and the upper-triangular entries of Λ (owing to symmetry), and place a uniform prior on the diagonal entries of Λ:

$$\pi(\Phi) = \left[\prod_{i,j} \pi_{\mathrm{SS}}(\Theta_{ij})\right] \times \left[\prod_{i<j} \pi_{\mathrm{SS}}(\Lambda_{ij})\right] \times \left[\prod_{i} \pi_{\mathrm{Unif}}(\Lambda_{ii})\right].$$

The support of the joint prior distribution is the set $\{(\Theta, \Lambda) : \Lambda \succ 0, \|\Lambda\|_2 \leq R\}$, where $\Lambda \succ 0$ means that Λ is positive definite. We constrain the matrix L_2-norm of Λ to be upper bounded. Although this additional side constraint adds a restriction to the high-dimensional parameter space, it is not that restrictive because the upper bound R is allowed to change with (n, p, q) and can be quite large.

2.3. MAP estimator: a penalized likelihood perspective

For computational efficiency, we estimate (Θ, Λ) using the posterior mode. The negative log posterior can be written as

$$L(\Phi) = -\ell(\Phi) + \sum_{i,j} \mathrm{pen}_{\mathrm{SS}}(\Theta_{ij}) + \sum_{i<j} \mathrm{pen}_{\mathrm{SS}}(\Lambda_{ij}), \tag{2.6}$$

where $\ell(\cdot)$ is the log-likelihood function (2.2), and $\mathrm{pen}_{\mathrm{SS}}(\cdot)$ is the negative logarithm of the spike and slab Lasso prior (2.5):

$$\mathrm{pen}_{\mathrm{SS}}(\theta) = -\log\left(\frac{\eta}{2v_1}e^{-|\theta|/v_1} + \frac{1-\eta}{2v_0}e^{-|\theta|/v_0}\right). \tag{2.7}$$

Finding the MAP estimator of (Θ, Λ) is equivalent to solving the optimization problem

$$\underset{\Theta, \Lambda \succ 0, \|\Lambda\|_2 \leq R}{\operatorname{argmin}} \; L(\Phi). \tag{2.8}$$

The minimizer of (2.6) has a natural interpretation as the penalized likelihood estimator using the penalty function (2.7), which is induced by the Bayesian spike and slab Lasso prior. In the penalized likelihood framework, the derivative of a penalty function $\mathrm{pen}'_{\mathrm{SS}}(\theta)$ often plays the role of thresholding. An ideal property of a penalty function is to threshold adaptively: $\mathrm{pen}'_{\mathrm{SS}}(\theta)$ is large when θ is small, so the resulting estimate is exactly zero, and $\mathrm{pen}'_{\mathrm{SS}}(\theta)$ is small when θ is large, so the resulting estimate is almost unbiased, without being affected by the thresholding value. It is well known that the Bayesian penalty induced from a single Laplace prior $\mathrm{LP}(\theta; v)$ is equivalent to the ℓ_1-penalty (Tibshirani

(1996); Park and Casella (2008)), the derivative of which takes a constant value, and therefore, does not possess such an adaptive property, which is particularly helpful for achieving structure recovery properties.

In the proposition below, which is a generalization of Lemma 1 of Ročková and George (2018), we show that the first and second derivatives of our Bayesian penalty function, induced by the spike and slab Lasso prior, can be linked to the mean and variance of a family of binary random variables.

Proposition 1. *$pen_{SS}(\theta)$ is a concave function when θ is in $\mathbb{R}^+$, with*

$$pen'_{SS}(\theta) = \mathbb{E}Z(\theta) = \frac{\eta(\theta)}{v_1} + \frac{1-\eta(\theta)}{v_0},$$
$$pen''_{SS}(\theta) = -Var(Z(\theta)) = \eta(\theta)(1-\eta(\theta))\left(\frac{1}{v_0} - \frac{1}{v_1}\right)^2,$$

where $Z(\theta)$ is a binary random variable taking the value $1/v_1$ with probability $\eta(\theta)$, and the value $1/v_0$ with probability $1-\eta(\theta)$, where $\eta(\theta)$ is given by $\eta(\theta) = \eta LP(\theta; v_1)/(\eta LP(\theta; v_1) + (1-\eta)LP(\theta; v_0))$.

A consequence of Proposition 1 is that the spike and slab Lasso prior leads to an adaptive regularization procedure: $\text{pen}'_{\text{SS}}(\theta)$ is a decreasing function with respect to the magnitude of θ. In particular, the penalty at θ is a weighted average of a large penalty $1/v_0$ and a small one $1/v_1$, where the weights $\eta(\theta)$ and $1-\eta(\theta)$ are the conditional probabilities of θ belonging to the "slab" and "spike" components, respectively.

3. Theoretical Results

For our theoretical studies, we evaluate the performance of our Bayesian procedure under the frequentist data-generating mechanism, that is, under the assumption that the data Y are generated based on a fixed set of parameters Φ^0. This is a common practice in theoretical analyses of Bayesian methods such as those of Ishwaran and Rao (2005), Narisetty and He (2014), Castillo, Schmidt-Hieber and Van der Vaart (2015), and Gan, Narisetty and Liang (2019).

We first provide the optimal ℓ_2-error bounds for all points from the HPD region,

$$\begin{aligned} \text{HPD} &= \{\Phi : \pi(\Phi \mid \text{Data}) \ \geq \ \pi(\Phi^0 \mid \text{Data})\} \\ &= \{\Phi : L(\Phi) \ \leq \ L(\Phi^0)\}, \end{aligned} \tag{3.1}$$

and show that there exists at least one local optimum in the HPD that has the

optimal error rate in the ℓ_∞-norm, and that it has the same support as the true graph. We further show that the optimal error rate in the ℓ_∞-norm holds for all local modes of the fractional posterior, that is, the posterior defined with respect to a fractional likelihood.

Our results for the optimal error rate in the ℓ_∞-norm lead to support recovery consistency, without the incoherence condition required by Wytock and Kolter (2013). Note that some existing works also do not require the incoherence condition, but there are important differences between the respective results. The results for the smoothly clipped absolute deviations (SCAD) penalty in Fan and Li (2001) are valid for only one of the local solutions, whereas our results ascertain the consistency for all solutions. The results of Cai, Liu and Luo (2011) are applicable only to unconditional graphical models, and are not directly applicable to settings with covariates that use the GCRF model, which is quite different from unconditional graphical models.

Notation. Denote the true parameters as Φ^0, Λ^0, and Θ^0. Let $S_0 = \{(i,j) : \Phi^0_{ij} \neq 0\}$ denote the signal set, $\theta^0_{\min} = \max_{(i,j)\in S_0} |\Phi^0_{ij}|$ be the minimal signal strength, and $d = \max_{i=1:(p+q)} \operatorname{card}\{j : \Phi^0_{ij} \neq 0\}$ be the maximum degree of the underlying conditional graph. We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the largest and smallest eigenvalues, respectively, of a symmetric matrix A, and $|||\cdot|||_\infty$ to denote the ℓ_∞/ℓ_∞ operator norm of a matrix. Define

$$c_{\Theta^0} = |||(\Theta^0)^T|||_\infty, \quad c_{\Lambda^0} = |||(\Lambda^0)^{-1}|||_\infty, \quad c_H = \frac{n}{2}|||H^{-1}_{S_0S_0}|||_\infty,$$

where $H := \nabla^2\ell(\Phi^0)$ denotes the Hessian matrix evaluated at Φ^0, and $H^{-1}_{S_0S_0}$ denotes a submatrix of H^{-1} with row and column indices from the set S_0. Note that H is a matrix with dimension equal to the total number of parameters in Φ. Furthermore, define $K_* = 8\max_i \Sigma^0_{ii} + 8\max_i \left((\Lambda^0)^{-1}\Theta^T\Sigma^0_{xx}\Theta(\Lambda^0)^{-1}\right)_{ii}$, where Σ^0 denotes the covariance matrix of (X,Y), and Σ^0_{xx} is the covariance matrix of X.

Note that the symbols used in our theorems and proofs do not represent fixed constants, and may vary with n, unless otherwise specified. We drop the subscript n from symbols; otherwise, we would write $p = p_n$, and $\Phi^0 = \Phi^0_n$. Our notation is similar to that used in Loh and Wainwright (2017). In particular, this implies that the minimum and maximum nonzero entries of the true parameters can depend on the sample size. Note that K_* is upper bounded by the maximum variance of all variables. Therefore, K_* can be upper bounded by a fixed constant not depending on n if the the variances of the covariates and the response variables

are upper bounded.

3.1. Preliminary results

Before presenting our results, we first present some preliminary results from the literature on the log-likelihood function $\ell(\Phi)$.

In our analysis, we examine $\ell(\Phi)$ in a small neighborhood around the true parameter value Φ^0. Expand $\nabla\ell(\Phi^0+\Delta)$ as follows:

$$\nabla\ell(\Phi^0) + H\cdot\mathrm{vec}(\Delta) + R(\Delta), \tag{3.2}$$

where $H=\nabla^2\ell(\Phi^0)$ is the Hessian matrix, and $R(\Delta)=\nabla\ell(\Phi^0+\Delta)-\nabla\ell(\Phi^0)-H\cdot\mathrm{vec}(\Delta)$ denotes the residual. The following lemma provides some useful bounds for $\nabla\ell(\Phi^0)$ and $R(\Delta)$. We omit the proof here, because the first bound is the same as Proposition 4 in Yuan and Zhang (2014), and the second bound is similar to Lemma 3 in Wytock and Kolter (2013), with their notation of $\|S_{xx}\|_\infty \le c_X^2$ replaced by $\|S_{xx}\|_\infty \le 9\rho_2$, where $\rho_2 = 1.5\lambda_{\max}(\Sigma^0_{xx})$. Note that our log-likelihood function $\ell(\Phi)$ differs from theirs by a factor $n/2$.

Lemma 1. *Assume data are generated from a GCRF model with true parameter* Φ^0.

1. *We have* $\|\nabla\ell(\Phi^0)\|_\infty \le K_*\sqrt{n\log(10(p+q)^2/\eta)}$ *with probability* $1-\epsilon_0$, *given the sample size* $n \ge \log(10(p+q)^2/\epsilon_0)$, *where* ϵ_0 *is any constant in* $(0,1)$.
2. *If* $\|\Delta\|_\infty \le (1/d)\min\{1/3c^0_\Lambda, c_{\Theta^0}/2\}$, *then* $(2/n)\|R(\Delta)\|_\infty \le 1{,}854d^2c^2_{\Theta^0}c^4_{\Lambda^0}\rho_2\|\Delta\|_\infty$, *where* $\rho_2 = 1.5\lambda_{\max}(\Sigma^0_{xx})$.

The local strong convexity of the log-likelihood function $\ell(\Phi)$ plays an important role in our theoretical analysis. Following Yuan and Zhang (2014), we define the local restricted strong convexity (LRSC) constant, a quantity that measures the local curvature of $\ell(\Phi)$ at Φ^0:

$$\beta(\Phi^0; r, \alpha) = \inf\left\{ \frac{\langle \nabla\ell(\Phi^0+\Delta) - \nabla\ell(\Phi^0), \Delta\rangle}{\|\Delta\|_2^2} : \right.$$
$$\left. \|\Delta\|_2 \le r, \|\Delta_{S_0^c}\|_1 \le \alpha\|\Delta_{S_0}\|_1 \right\}.$$

We next state an assumption that is needed in our theoretical analysis.

Assumption 1. *Assume that the covariate vector* X *is from a random design with covariance matrix* Σ^0_{xx}, *and satisfies the following* s_0*-sparse restricted isom-*

etry property condition:

$$\begin{cases} \inf\left(\dfrac{u^T S_{xx} u}{u^T \Sigma^0_{xx} u} : u \neq 0, \|u\|_0 \leq s_0\right) \geq 0.5, \\ \sup\left(\dfrac{u^T S_{xx} u}{u^T \Sigma^0_{xx} u} : u \neq 0, \|u\|_0 \leq s_0\right) \leq 1.5, \\ \dfrac{\lambda_{\max}[(\Theta^0)^T S_{xx} \Theta^0]}{\lambda_{\max}[(\Theta^0)^T \Sigma^0_{xx} \Theta^0]} \leq 1.4. \end{cases}$$

*The same assumption is made by Yuan and Zhang (*2014*) to analyze the GCRF model with the ℓ_1-penalty, and is also frequently used in compressed sensing. It is well known that this condition holds with high probability when X is sub-Gaussian, with a well conditioned population covariance matrix satisfying certain regularity assumptions on the eigenvalues, and n sufficiently large, for example, $n \geq O\big((p+s_0)\log(p+q)\big)$ (Candes and Tao (*2007*); Yuan and Zhang (*2014*)).*

The following lemma, which summarizes Proposition 3 from Yuan and Zhang (2014), ensures that $\beta(\Phi^0; r, \alpha)$ is positive for a GCRF model with high probability when the sample size n is sufficiently large. That is, $\ell(\Phi)$ behaves like a strongly convex function locally in the cone $\|\Delta_{S_0^c}\|_1 \leq \alpha\|\Delta_{S_0}\|_1$, although $\ell(\Phi)$ is not a strongly convex function at Φ^0.

Lemma 2. *Let*

$$\begin{aligned} &\rho_1 = 0.5 \min\left(\lambda_{\max}(\Lambda^0)^{-1}\lambda_{\min}(\Sigma^0_{xx})\right), \quad \rho_2 = 1.5\lambda_{\max}(\Sigma^0_{xx}), \\ &r_0 = \min\left[0.5\lambda_{\min}(\Lambda^0),\ 0.13\sqrt{\frac{\lambda_{\max}[(\Theta^0)^T\Sigma^0_{xx}\Theta^0]}{\rho_2}}\right], \\ &\beta_0 = \left\{\frac{\rho_1}{40\lambda_{\max}(\Lambda^0)} \cdot \min\left[1, \frac{\lambda_{\min}(3\Lambda^0)}{16\lambda_{\max}\left((\Theta^0)^T\Sigma^0_{xx}\Theta^0\right)}\right]\right\}. \end{aligned}$$

Assume that Assumption 1 *holds with*

$$s_0 = |S_0| + \left\lceil 4\left(\frac{\rho_2}{\rho_1}\right)\alpha^2|S_0|\right\rceil. \tag{3.3}$$

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

3.2. Rate of convergence for all points in the HPD

We first show that for any point Φ in the HPD region, its error term $\Delta = \Phi - \Phi^0$ belongs to a cone if $1/v_1$ is chosen properly.

Lemma 3. *If* $1/v_1 > 2\|\ell(\Phi^0)\|_\infty$, *then for any* $\Phi = \Phi^0 + \Delta$ *such that* $L(\Phi) \leq L(\Phi^0)$, *we have* $\|\Delta_{S_0^c}\|_1 \leq \alpha\|\Delta_{S_0}\|_1$, *where* $\alpha = 1 + 2v_1/v_0$.

We then show that all points from the HPD region, including the global maximum and all stationary points of the posterior distribution with $L(\Phi) \leq L(\Phi^0)$, are close to the true parameter value within an optimal statistical precision. Our analysis allows the quantities (v_0, v_1, R) and the model size p, q, and $|S_0|$ to grow with the sample size n. However, we suppress this dependence on n in our notation, for convenience.

Theorem 1. *(Rate of convergence for all points in the HPD). Assume Assumption* 1 *holds, with* s_0 *defined in* (3.3). *If*

(i) the prior hyperparameters v_0, *and* v_1 *satisfy*

$$\frac{2\|\nabla\ell(\Phi^0)\|_\infty}{n} \leq \frac{1}{nv_1} = C_1\sqrt{\frac{\log(p+q)}{n}}, \quad \frac{1}{nv_0} = C_0\sqrt{\frac{\log(p+q)}{n}},$$

for some constants $C_0 \geq C_1$,

(ii) the matrix norm bound R *satisfies* $R < 2\lambda_{\min}(\Lambda^0)\sqrt{r_0}/\varepsilon_n$, *and*

(iii) the sample size n *satisfies* $n \geq \log(10(p+q)^2/\epsilon_0)$,

then for any Φ *from the HPD region* (3.1), *we have*

$$\|\Phi - \Phi^0\|_F \leq \varepsilon_n := \frac{C_0 + C_1}{\beta_0}\sqrt{\frac{|S_0|\log(p+q)}{n}}$$

with probability no less than $1 - \epsilon_0$, *where* ϵ_0 *is a constant from* $(0, 1)$.

Our conditions and theoretical results require the following condition on the magnitude of the relationship among (p, q, n): $(p + s_0)\log(p + q) = o(n)$. This ensures that Assumption 1 holds with high probability, and that the F-norm estimation error bound of Theorem 1 goes to zero. This is not a restrictive requirement because our focus is high-dimensional settings where $\dim(X) = q \gg \dim(Y) = p$, and it still allows the covariate dimension q to be much larger than the sample size.

A proof of Theorem 1 is provided in the Supplementary Material. Our proof is motivated by Theorem 1 of Yuan and Zhang (2014). However, the proof technique in Yuan and Zhang (2014) is tailored to the Lasso penalty, which needs to be extended to handle our concave penalty function, $\text{pen}_{\text{SS}}(\theta)$.

Theorem 1 does not impose any conditions on the mixing weight η and the difference between the two scale parameters v_0 and v_1. Therefore, Theorem 1

includes special cases such as $\eta = 0$, $\eta = 1$, and $v_1 = v_0$. In these cases, the spike and slab Lasso penalty degenerates to the ordinary Lasso penalty with one unique global optimum. For the Lasso penalty, Yuan and Zhang (2014) established a similar error bound for the global optimum. However, our result is stronger because it establishes the error bound for all points in the HPD, including the global optimum.

3.3. Faster rate of convergence for a local optimum and its sparsistency

The result in Section 3.2 is for all points in the HPD region. Next, we provide stronger results for the estimation and selection accuracy for at least one local optimum in the HPD.

Theorem 2 (Rate of convergence in the ℓ_∞-norm and sparsistency)**.** *Assume Assumption* 1 *holds, with* s_0 *defined in* (3.3)*. Then, there exists a stationary point* $\tilde{\Phi}$ *in the HPD, such that*

$$\tilde{\Phi}_{S_0^c} = 0, \quad \|\tilde{\Phi} - \Phi^0\|_\infty \leq r_n := 4c_H(C_1 + C_0)\sqrt{\frac{\log(p+q)}{n}},$$

with probability $1 - \epsilon_0$*, if the following conditions hold:*

(i) the prior hyperparameters v_0, v_1*, and* η *satisfy* $0 < \eta \sim O(1) < 1$*,*

$$\frac{2\|\nabla \ell(\Phi^0)\|_\infty}{n} \leq \frac{1}{nv_1} = C_1\sqrt{\frac{\log(p+q)}{n}}, \quad \frac{1}{nv_0} = C_0\sqrt{\frac{\log(p+q)}{n}},$$

for some constants $C_0 > C_1$*;*

(ii) $\theta^0_{\min} > r_n + \delta_0$*, where* $\delta_0 > [n\log(p+q)]^{-\alpha/2}$*, with* $0 < \alpha < 1$ *and*

$$r_n \leq \min\left\{\frac{1}{3c_{\Sigma^0}d}, \frac{1}{3{,}708d^2c^2_{\Gamma^0}c^4_{\Sigma^0}\rho_2}, \frac{c_{\Theta^0}}{2d}, \frac{2pen'_{SS}(0^+)}{n(c_H + 1/c_H)}\right\};$$

(iii) the sample size n *satisfies* $n \geq \log(10(p+q)^2/\epsilon_0)$*.*

The condition $\theta^0_{\min} > r_n + \delta_0$ is the usual beta-min condition, meaning that the minimal signal strength in Φ_{S_0} should be bigger than the ℓ_∞-error bound by a small margin δ_0, where δ_0 can go to zero at a rate slower than $[n\log(p+q)]^{-\alpha}$. Under the beta-min condition, Theorem 2 ensures that $\min_{(i,j)\in S_0}|\tilde{\Phi}_{ij}| \geq \delta_0$. Consequently, $\tilde{\Phi}$ achieves sparsistency; that is, $\tilde{\Phi}_{S_0^c} = 0$ and $\tilde{\Phi}_{S_0} \neq 0$.

In contrast to Theorem 1, Theorem 2 requires η to be strictly between zero and one, and v_1 to be strictly bigger than v_0; thus the ordinary one-component

Laplace prior, that is, the Lasso penalty, does not satisfy the assumptions here. Note that our theoretical results do not require that η be small and decrease to zero with the dimension, nor do they allow v_1/v_0 to diverge. This appears to contradict prior results on Bayesian variable selection using spike and slab priors, such as those of George and McCulloch (1993), Ishwaran and Rao (2005), Narisetty and He (2014), and Castillo, Schmidt-Hieber and Van der Vaart (2015). The main reason for this difference is that these approaches consider the integrated posterior on all models, after integrating out the continuous model parameters owing to which they require a multiplicity adjustment for a large number of models. In contrast, because our theoretical analysis studies the posterior on the continuous model parameters directly, our conditions on the prior parameters v_1 and η do not have a direct correspondence with the previous choices. In particular, our theoretical results are under the condition that v_1 is not much larger than v_0, because a larger gap between them would imply more nonconvexity of the negative log posterior, which makes it difficult to compute and theoretically study its stationary points.

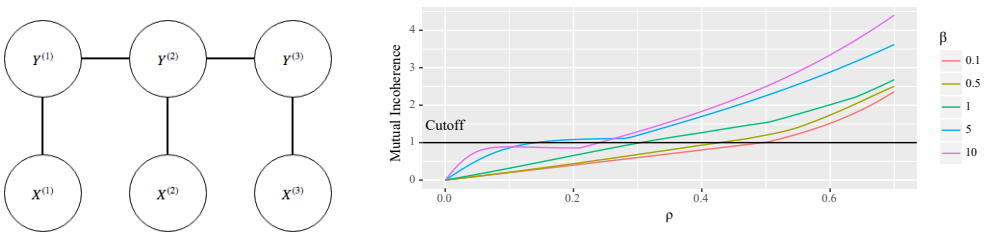
A proof for Theorem 2 is provided in the Supplementary Material, which is motivated by similar results by Ravikumar et al. (2011), Wytock and Kolter (2013), Loh and Wainwright (2017), and Gan, Narisetty and Liang (2019). We start with a restricted optimization problem

$$\min_{\Lambda \succ 0, \Phi_{S_0^c}=0} L(\Phi), \tag{3.4}$$

and then show that there exists a solution $\tilde{\Phi}$ to (3.4) that satisfies $\|\tilde{\Phi}-\Phi^0\|_\infty \leq r_n$. The last and most important step is to prove that $\tilde{\Phi}$ is indeed 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 $\tilde{\Phi}$.

Previously, under mutual incoherence conditions, Wytock and Kolter (2013) showed that the convergence rate in the ℓ_∞-norm for the GCRF model with ℓ_1- penalty is of the same order as ours. However, their approach requires the restrictive mutual incoherence condition, that is, $\|\|H_{S_0^c S_0}(H_{S_0 S_0})^{-1}\|\|_\infty < 1$, which our approach does not require. We illustrate that this condition can be easily violated in the following toy example. Consider a simple Markov chain GCRF model in Figure 1(a), with

$$\Lambda^0 = \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix}, \quad \Theta^0 = \begin{bmatrix} \rho\beta & 0 & 0 \\ 0 & \rho\beta & 0 \\ 0 & 0 & \rho\beta \end{bmatrix}.$$



(a) Markov chain GCRF model (b) Mutual incoherence is violated for large ρ

Figure 1. Violation of mutual incoherence condition for the chain graph.

In Figure 1(b), we plot $\|\|H_{S_0^c S_0}(H_{S_0 S_0})^{-1}\|\|_\infty$ for five different choices of β. For each β, the mutual incoherence condition is violated once ρ is larger than some threshold.

3.4. On the uniqueness of the stationary points in the HPD

Although Theorem 2 asserts that there is one stationary point in the HPD region that has a desired rate of convergence, it is natural to ask whether the stationary point is unique. Unfortunately, we cannot ascertain this directly for the posterior distribution using the spike and slab regularization. However, if we consider a slightly modified version of the negative log-posterior minimization (2.6) given by

$$L_\kappa(\Theta) = -\ell(\Phi) + \kappa \mathrm{Pen}(\Phi), \tag{3.5}$$

where κ is a parameter that enhances the amount of Bayesian regularization, the stationary solution can be proved to be unique for some choices of κ.

The modified objective function (3.5) can be viewed as the negative log-posterior corresponding to the fractional posterior distribution $\pi_\kappa(\Theta \mid \text{Data})$, which is the posterior distribution defined with respect to the likelihood of the data raised to the power $1/\kappa$, that is, $\pi_\kappa(\Theta \mid \text{Data}) \propto \exp(\ell(\Phi)/\kappa - \mathrm{Pen}(\Phi))$. The HPD region corresponding to this fractional posterior distribution can be defined accordingly as

$$\{\Phi : L_\kappa(\Phi) \leq L_\kappa(\Phi^0)\}. \tag{3.6}$$

Next, we show that Theorem 1 and Theorem 2 can be extended to cover the fractional posterior. In addition, we can show that with a proper choice of hyperparameters, the HPD region is unimodal, with a unique stationary point that achieves the desired ℓ_∞-accuracy.

Theorem 3. *Assume Assumption* 1 *holds, with* s_0 *defined in* (3.3). *Further, assume the following conditions hold:*

(i) $\kappa = \log(p+q)$;

(ii) *the prior hyperparameters* v_0, v_1, *and* η *satisfy* $0 < \eta \sim O(1) < 1$,

$$\frac{\|\nabla \ell(\Phi^0)\|_\infty}{n} \leq \frac{\kappa}{nv_1} = C_1\sqrt{\frac{\log(p+q)}{n}}, \quad \frac{\kappa}{nv_0} = C_0\sqrt{\frac{\log(p+q)}{n}},$$

for some constants $C_0 > C_1$;

(iii) *the matrix norm bound* R *satisfies* $R < 2\lambda_{\min}(\Lambda^0)\sqrt{r_0}/\varepsilon_n$; *and*

(iv) *the sample size* n *satisfies* $n \geq \log(10(p+q)^2/\epsilon_0)$. *Then, with probability going to one, for any point in the HPD region* (3.6), *we have*

$$\|\Phi - \Phi^0\|_F \leq \varepsilon_n := \frac{C_0 + C_1}{\beta_0}\sqrt{\frac{|S_0|\log(p+q)}{n}}.$$

Further, if we assume r_n *satisfies* $\theta^0_{\min} - r_n > \delta_0$, *where* $\delta_0 > [n\log(p+q)]^{-\alpha/2}$, *with* $0 < \alpha < 1$, *and*

$$r_n \leq \min\left\{\frac{1}{3c_{\Sigma^0}d}, \frac{1}{3{,}708d^2c^2_{\Gamma^0}c^4_{\Sigma^0}\rho_2}, \frac{c_{\Theta^0}}{2d}, \frac{2pen'_{SS}(0^+)}{n(c_H + 1/c_H)}\right\},$$

then the unique stationary point $\tilde{\Phi}$ *from the HPD region satisfies*

$$\tilde{\Phi}_{S_0^c} = 0, \quad \|\tilde{\Phi} - \Phi^0\|_\infty \leq r_n := 4c_H(C_1 + C_0)\sqrt{\frac{\log(p+q)}{n}},$$

with probability converging to 1.

4. Empirical Results

4.1. Simulation studies

In the simulation studies, we compare different methods in terms of their parameter estimation, structure recovery, and prediction. Following Yuan and Zhang (2014), we generate X from a zero-mean multivariate Gaussian distribution with a *dense* precision matrix $\Theta^0_{xx} = 0.5(J + I)$, where J is a matrix of ones, and generate Y given X from the GCRF model (2.1), with the true (Θ^0, Λ^0) generated as follows. The precision matrix Λ^0 is generated as a random graph, similar to the setup of the random graph in Peng et al. (2009). We first generate the entries in the precision matrix following the distribution of $S \times B \times U_1$, where $(S+1)/2 \sim \mathsf{Bern}(0.5)$, $B \sim \mathsf{Bern}(0.1)$, $U_1 \sim \text{Uniform}(1, 2)$, and the three random variables are independent. We then rescale the nonzero elements to ensure the

positive definiteness of Λ. Specifically, we first sum the absolute value of each row, and then divide each off-diagonal entry by 1.1 fold of it. We then average the rescaled matrix with its transpose to ensure symmetry. Finally, the diagonal entries are all set to one. We consider the following forms of true Θ^0:

1. Model 1 (Random Graph): entries in Θ^0 are generated as $S \times B \times U_2$ where S and B are random variables, as defined before, and are independent of $U_2 \sim \text{Uniform}(0.5, 1)$.

2. Model 2 (Banded Model 1): for the ith row of Θ^0, the $((i-1)/\lfloor q/p \rfloor + 1)$th element is generated from $S \times B \times U_2$. All other entries are zero.

3. Model 3 (Banded Model 2): the ith row of Θ^0 has probability 0.1 of being nonzero and probability 0.9 of being all zero; when the ith row of the Θ^0 is nonzero, its entries are generated from the distribution of $S \times B \times U_2$, where $(S+1)/2 \sim \text{Bern}(0.5)$, $B \sim \text{Bern}(0.1)$, and $U_2 \sim \text{Uniform}(0.5, 1)$.

For each model, we fix the observation size as $n = 100$ and the dimension of the outcome vector as $p = 50$, and take the covariate dimension q to be $(50, 100, 200, 500)$. The results are summarized based on 100 replications. We report three metrics to measure the estimation, selection, and prediction accuracy of each method: *i*) for the estimation accuracy, we use the Frobenius norm distance (denoted as Fnorm); *ii*) for the selection accuracy of the structure recovery, we use the Matthews correlation coefficient (MCC):

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}},$$

where TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively; *iii*) for the prediction accuracy, we use the average MSE on an independently generated test data set of size 100. Note that it may not be meaningful to compare results across different values of q, because the level of sparsity in (Θ^0, Λ^0) and the magnitude of the signal in Λ change with q. Thus, we recommend comparing the results across different methods for the same value of q.

In the simulation studies, we compare our method, denoted as BayesCRF, with the following alternative methods: 1)the GCRF model with ℓ_1-regularization, based on the implementation of Wytock and Kolter (2013), and denoted as L1-GCRF; 2) a joint graphical Lasso (Friedman, Hastie and Tibshirani (2008)) for (X, Y), denoted as GLasso; and 3) a covariate-adjusted graphical model proposed by Cai et al. (2012), denoted as CAPME. Because CAPME does not directly estimate Θ, we first estimate B, the regression coefficient matrix, and then use

Table 1. 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)**

the relationship given by (2.4) to recover Θ. We fix $v_0 = \sqrt{1/(n\log(p+q))}$, $v_1 = 3v_0$, and $\eta = 0.5$ for our BayesCRF method, with $\alpha = 1$ corresponding to the complete posterior. We choose the tuning parameters for the aforementioned alternatives using cross-validation, as suggested in the respective papers.

The results for the banded Model 1 are provided in Table 1. The results for the other models are presented in the Supplementary Material owing to space limitations but we comment on them here. We have the following conclusions from the results: 1) Our BayesCRF method achieves the best performance for parameter estimation (based on Fnorm), support recovery (based on MCC), and prediction (based on Test Error) in most of the cases considered. These results can be attributed to the adaptiveness of the spike and slab Lasso penalty. 2) The performance of GLasso is not as desirable as that of BayesCRF, likely because of the accumulation of errors in estimating the structure of X, which is not relevant to the parameters of the GCRF model, as discussed in Section 2. 3) CAPME exhibits poor performance in terms of the MCC and Fnorm measures, because it is not designed to detect the conditional dependence structure of interest. However, it works well for prediction because this depends on B alone. 4) L1-GCRF performs worse than BayesCRF, but performs better than the other competing methods in general, although its test error is too large in the random graph setting with $q = 500$ and $n = 100$.

4.2. Application: asset returns prediction

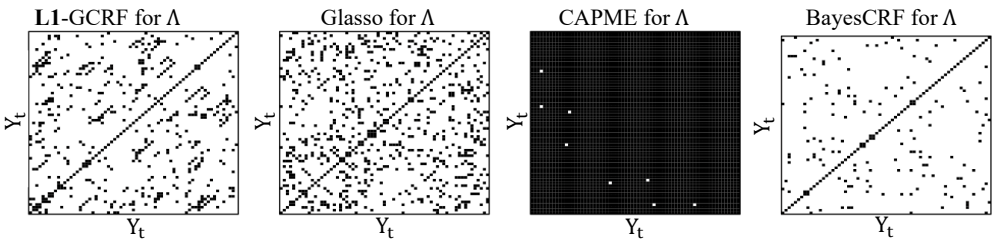
We now compare the performance of our method with that of the other alternatives for the problem of predicting asset returns. The data set we consider is the weekly price data of S&P 500 stocks for 265 consecutive weeks from March 10, 2003, to March, 24, 2008, collected by Pfaff (2016). We screen out all stocks with extremely low or high marginal variance, and keep 67 stocks that vary modestly, that is, stocks with a variance between 25 and 40. All stock prices are log transformed. Let $Y_t = [Y_t^1, \ldots, Y_t^{67}] \in \mathbb{R}^{67}$ denote the stocks prices at time point t, and let $X_t = [Y_{t-5}, Y_{t-4}, Y_{t-3}, Y_{t-2}, Y_{t-1}]$ denote the prices for the previous five weeks. We want to recover the dependence structure between Y_t and X_t, and within Y_t, to provide insights into the dependency between the prices of different stocks and between their previous prices. We also measure how well we can predict Y_t using X_t because we cannot directly evaluate the quality of the estimated structure.

We apply all methods to the first 212 days to estimate Φ, and make predictions on the remaining 53 days using equation (2.4). We first standardize all variables to have zero mean and unit variance. We then transform the data back to the original log-scale to make predictions. The tuning parameters for all methods are selected using five-fold cross-validation, and the average prediction errors are evaluated using

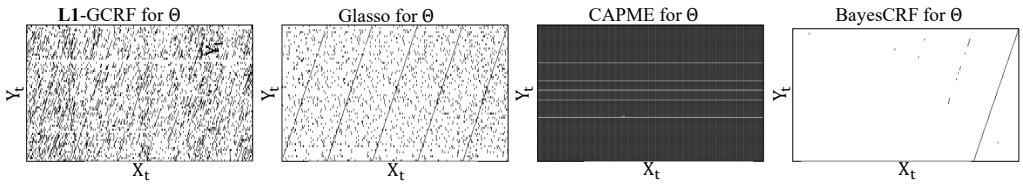
$$\overline{Err} = \frac{1}{49} \sum_{t=213}^{265} ||Y_t - \hat{Y}_t||_2.$$

The average prediction errors for the methods are provided in Table 2. BayesCRF achieves the smallest average prediction error. The prediction performance of GLasso and CAPME are similar, while the algorithm for L1-GCRF fails to make an accurate prediction.

The conditional graphs estimated from the methods are shown in Figure 2. We observe the following: 1) BayesCRF detects that some of the concurrent prices of the assets are conditionally dependent on each other (shown in the estimated Λ matrix), and there is an $AR(2)$-like structure for each asset over time (shown in the estimated Θ), that is, Y_t^i is conditionally dependent on Y_{t-1}^i, Y_{t-2}^i. GLasso and L1-GCRF detect much noisier patterns with longer time dependences. 2) BayesCRF provides sparser estimates of the matrices (Θ, Λ) and, at the same time, the best prediction accuracy. This suggests that BayesCRF provides a desirable estimation with both sparsity and accuracy. In practice, it is favorable to have sparser estimates because sparse models reduce the cost of data processing



(a) Estimates for the precision matrix Λ for the asset return data.



(b) Estimates for Θ for the asset return data. The ith horizontal axis tick (from left to right) represents the ith entry X_t, and the ith vertical axis tick (from down to top) represents the ith entry Y_t.

Figure 2. Estimates of the graphs in the asset returns application. White represents the noise, and black represents the selected signal.

Table 2. Average Prediction Error for Asset Return Prediction

BayesCRF	L1-GCRF	CAPME	Glasso
0.910(0.384)	3.817(0.468)	1.443(0.442)	1.250(0.495)

and management.

Supplementary Material

The online Supplementary Material provides details on the properties of the proposed Bayesian regularization function, the log likelihood function, our proposed EM algorithm for computations, and its derivation, proofs for all the technical results, and additional simulation results.

Acknowledgments

Naveen Narisetty gratefully acknowledges funding support from NSF grants DMS-1811768 and CAREER-1943500.

References

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9**, 485–516.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–2604.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106**, 594–607.

Cai, T. T., Li, H., Liu, W. and Xie, J. (2012). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100**, 139–156.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–2351.

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

Deshpande, S. K., Ročková, V. and George, E. I. (2017). Simultaneous variable and covariance selection with the multivariate spike-and-slab Lasso. *arXiv preprint arXiv:1708.08911*.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9**, 432–441.

Frot, B., Jostins, L. and McVean, G. (2019). Graphical model selection for Gaussian conditional random fields in the presence of latent variables. *J. Amer. Statist. Assoc.* **114**, 723–734.

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

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881–889.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33**, 730–773.

Lafferty, J., McCallum, A. and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning, ICML*, 282–289.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37**, 4254–4278.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16**, 559–616.

Loh, P.-L. and Wainwright, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *Ann. Statist.* **45**, 2455–2482.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436–1462.

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

Negahban, S., Yu, B., Wainwright, M. J. and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *NIPS*, 1348–1356.

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. *J. Amer. Statist. Assoc.* **104**, 735–746.

Pfaff, B. (2016). *Financial Risk Modelling and Portfolio Optimisation with R*. 2nd Edition. John Wiley & Sons, Ltd, London.

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

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

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

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

Ročková, V. and George, E. I. (2018). The spike-and-slab Lasso. *J. Amer. Statist. Assoc.* **113**, 431–444.

Rothman, A. J., Levina, E. and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Statist.* **19**, 947–962.

Sohn, K.-A. and Kim, S. (2012). Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *AISTATS*, 1081–1089.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58**, 267–288.

Witten, D. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high-dimensional problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71**, 615–636.

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

Yin, J. and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5**, 2630.

Yuan, X.-T. and Zhang, T. (2014). Partial Gaussian graphical model estimation. *IEEE Trans. Inf. Theory* **60**, 1673–1687.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27**, 576–593.

Lingrui Gan

Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA.

E-mail: lgan6@illinois.edu

Naveen N. Narisetty

Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA.

E-mail: naveen@illinois.edu

Feng Liang

Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA.

E-mail: liangf@illinois.edu

(Received September 2019; accepted June 2020)