Scalable Bayesian high-dimensional local dependence learning

이경재

인하대학교 통계학과

Joint work with

Lizhen Lin, The University of Notre Dame

Contents

Introduction

Estimation of covariance matrix

- The estimation of covariance (or its inverse) matrices is crucial to reveal the dependence structure.
- Many statistical methods require the estimated covariance matrix as the starting point of the analysis (e.g. LDA and QDA).
- Estimation of the covariance matrix is one of important and challenging tasks, especially when the number of variables is much larger than the sample size ($p \gg n$).

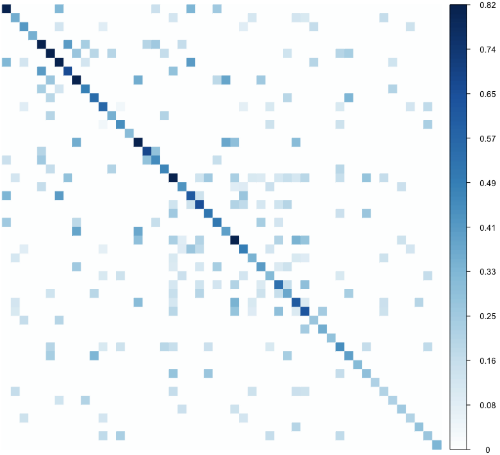


Figure: A sparse matrix.

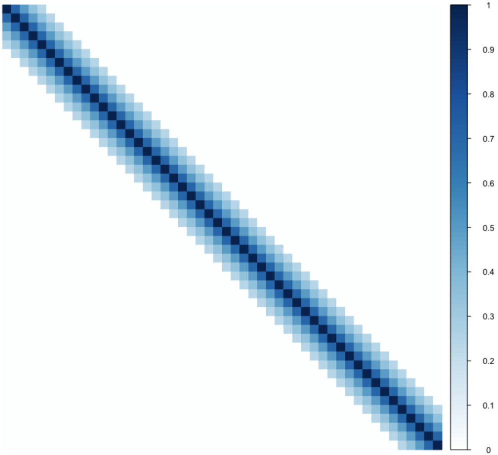


Figure: A banded matrix.

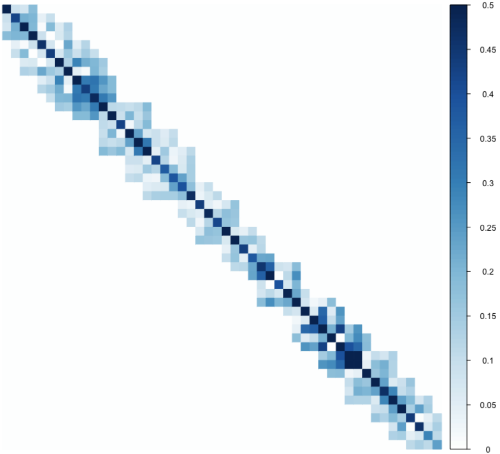


Figure: A banded matrix with varying bandwidths.

Restrictive matrix classes

- Various constraints can be encoded in many different ways which lead to different graph models:
  - on covariance matrices (Σ_n),
  - on precision matrices ($\Omega_n = \Sigma_n^{-1}$) or
  - on Cholesky factors (A_n, where $\Omega_n = (I_p - A_n)^T D_n^{-1} (I_p - A_n)$).
- In this talk, we focus Cholesky factors with varying bandwidths, which corresponds to a directed acyclic graph (DAG) model.

Main goals

$$X_1, \ldots, X_n \mid \Omega_n \overset{i.i.d.}{\sim} N_p\big(0, \Omega_n^{-1}\big), \quad \text{where } \Omega_n = (I_p - A_n)^T D_n^{-1} (I_p - A_n).$$

- Assume the high-dimensional settings, where $p \geq n$.
- Assume that the Cholesky factor A_n has varying bandwidths.
- The main goal is to develop a computationally scalable and theoretically sound Bayesian method.

Preliminaries

Modified Cholesky decomposition (MCD)

- (Modified Cholesky decomposition)

  For any positive definite matrix Ω_n, there exist unique

  - lower triangular matrix $A_n = (a_{jl})$ (Cholesky factor) and
  - diagonal matrix $D_n = diag(d_j)$ such that

  $$\Omega_n = (I_p - A_n)^T D_n^{-1} (I_p - A_n).$$

- Let $\Omega_n = \Sigma_n^{-1}$ be a $p \times p$ precision matrix.
- We assume a Cholesky factor with varying bandwidths.

Advantages

- The MCD-based approach has two advantages.
  - Positive definiteness of $\Omega_n = (I_p - A_n)^T D_n^{-1}(I_p - A_n)$.
  - A sequence of regression models interpretation:

$$Y = (Y_1, \ldots, Y_p)^T \mid \Omega_n \quad \sim \quad N_p(0, \Omega_n^{-1})$$

$$\Longleftrightarrow \quad \begin{cases} Y_1 \mid d_1 \sim N(0, d_1), \\ Y_j \mid Y_{1:(j-1)}, a_j, d_j \sim N\Big(\sum_{l=1}^{j-1} a_{jl} Y_l,\ d_j \Big),\ j = 2, \ldots, p \end{cases}$$

where $a_j = (a_{j1}, \ldots, a_{j,j-1})^T$ is the first $j-1$ elements of the jth row of A_n.

Banded Cholesky factor

- We assume a banded Cholesky factor A_n with varying bandwidths.
- If $Y \sim N_p(0, \Omega_n^{-1})$ with $\Omega_n = (I_p - A_n)^T D_n^{-1} (I_p - A_n)$ and the bandwidth of the jth row of A_n is k_j, it implies that

$$Y_j = a_{j,j-k_j} Y_{j-k_j} + \cdots + a_{j,j-1} Y_{j-1} + \epsilon_j, \tag{1}$$

  where $\epsilon_j \overset{ind}{\sim} N(0, d_j)$, for any $j = 2, \ldots, p$.
- We call (1) a local dependence structure.

Literature review

- Sparse Cholesky
  - Penalized likelihood approaches (Rothman et al. 2008, van de Geer and Bühlmann 2013, Khare et al. 2019)
  - Bayesian methods (Cao et al. 2019, Lee et al. 2019)
- Banded Cholesky with common bandwidth
  - Consistent test (An et al. 2014)
  - Bayesian methods (Banerjee and Ghosal 2014, Lee and Lee 2017, Lee and Lin 2020)
- Banded Cholesky with varying bandwidths
  - Penalized likelihood approach (Yu and Bien 2017)
  - No Bayesian method available

Bayesian local dependence learning

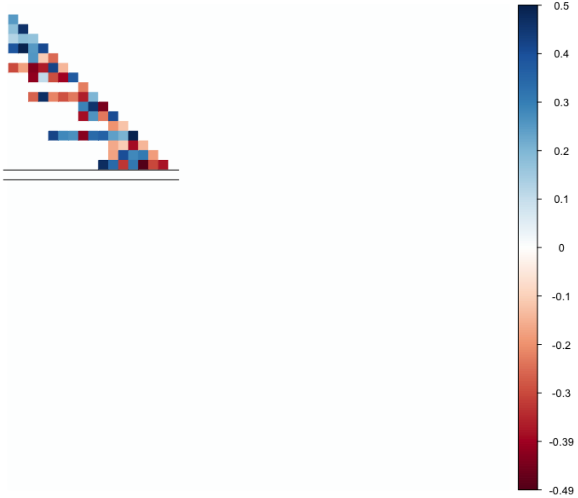
Gaussian Model with MCD

$$X_1, \dots, X_n \mid \Omega_n \overset{i.i.d.}{\sim} N_p(0, \Omega_n^{-1}), \quad \Omega_n = (I_p - A_n)^T D_n^{-1} (I_p - A_n)$$

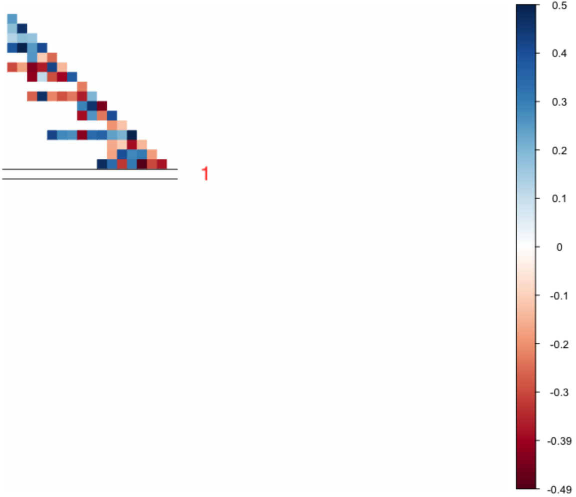
$$\Longleftrightarrow \quad \begin{cases} \tilde{X}_1 \mid d_1 \sim N_n\big(0,\ d_1 I_n\big), \\ \tilde{X}_j \mid \mathbf{X}_{j(k_j)}, a_j^{(k_j)}, d_j, k_j \overset{ind.}{\sim} N_n\Big(\mathbf{X}_{j(k_j)} a_j^{(k_j)},\ d_j I_n\Big), \quad j = 2, \dots, p \end{cases}$$

- $\tilde{X}_j \in \mathbb{R}^n$ is the jth column of the data matrix $\mathbf{X}_n = (X_1, \dots, X_n)^T \in \mathbb{R}^{n \times p}$.
- $\mathbf{X}_{j(k_j)} \in \mathbb{R}^{n \times k_j}$ is a submatrix of $\mathbf{X}_n$ consisting of the $(j - k_j), \dots, (j-1)$th columns of $\mathbf{X}_n$.
- $a_j^{(k_j)} \in \mathbb{R}^{k_j}$ is the $(j - k_j), \dots, (j-1)$th elements of $a_j = (a_{j1}, \dots, a_{j,j-1})^T$.

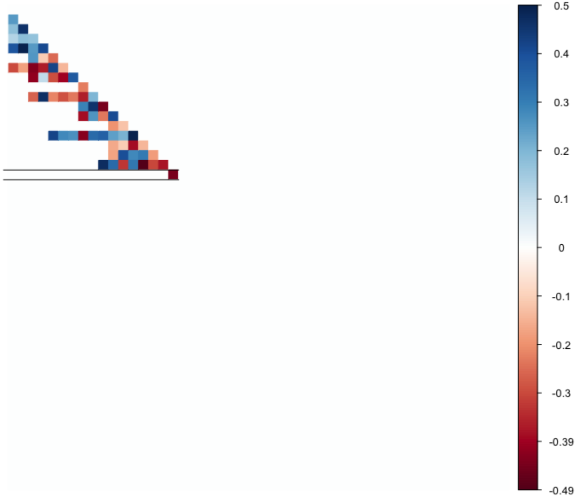
Prior for the Cholesky factor A_n (Sketch)



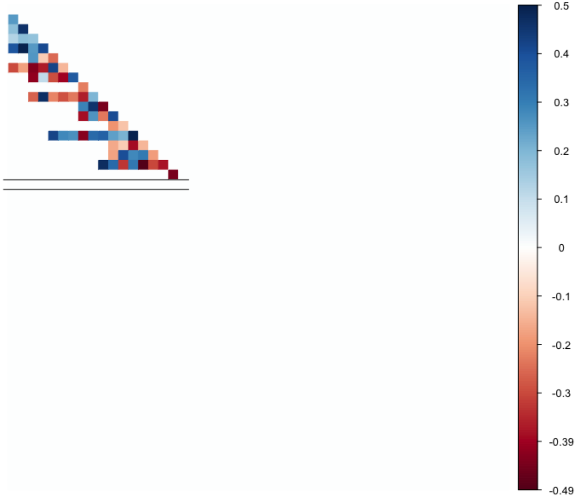
Prior for the Cholesky factor A_n (Sketch)



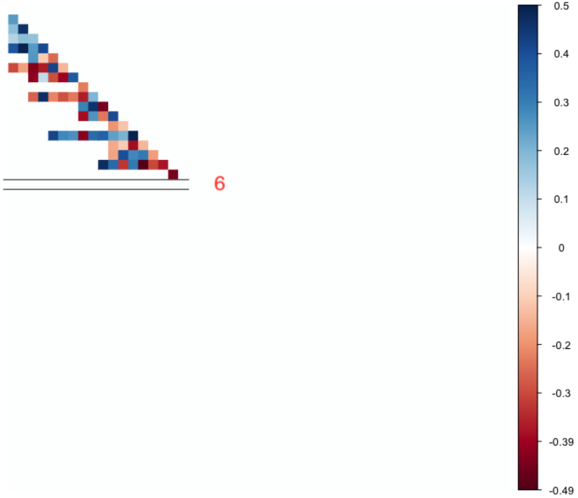
Prior for the Cholesky factor A_n (Sketch)



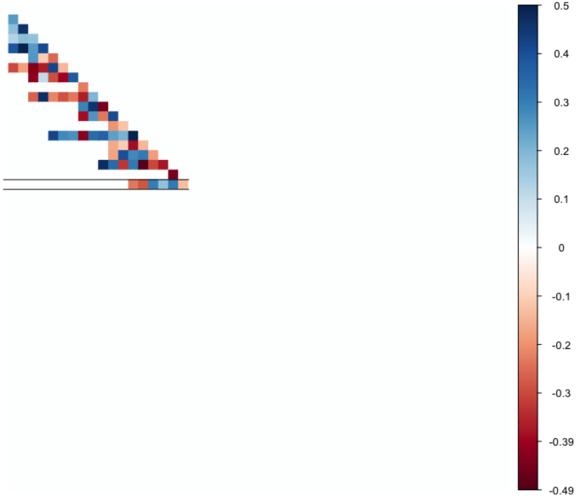
Prior for the Cholesky factor A_n (Sketch)



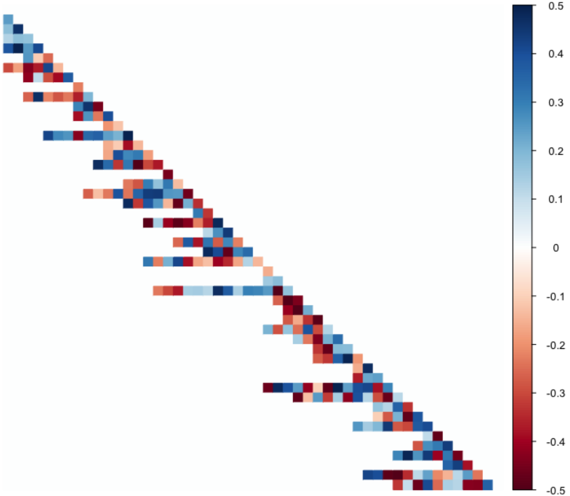
Prior for the Cholesky factor A_n (Sketch)



Prior for the Cholesky factor A_n (Sketch)



Prior for the Cholesky factor A_n (Sketch)



LANCE (LocAl depeNdence CholEsky) prior

$$\pi(A_n, D_n) = \pi(d_1) \prod_{j=2}^{p} \pi(a_j^{(k_j)} \mid d_j, k_j)\, \pi_j(k_j)\, \pi(d_j)$$

(i) Impose priors for k_j's:

$$\pi_j(k_j) \propto p^{-c_1 k_j} \quad \text{for } 0 \le k_j \le \{R_j \wedge (j-1)\}.$$

LANCE (LocAl depeNdence CholEsky) prior

$$\pi(A_n, D_n) = \pi(d_1) \prod_{j=2}^{p} \pi(a_j^{(k_j)} \mid d_j, k_j)\, \pi_j(k_j)\, \pi(d_j)$$

(ii) For the nonzero elements in a_j, impose a version of the Zellner's g-prior:

$$a_j^{(k_j)} \mid d_j, k_j \overset{ind.}{\sim} N_{k_j}\left(\widehat{a}_j^{(k_j)},\ \frac{d_j}{\gamma}\left(\mathbf{X}_{j(k_j)}^T \mathbf{X}_{j(k_j)}\right)^{-1}\right),$$

where $\widehat{a}_j^{(k_j)} = \left(\mathbf{X}_{j(k_j)}^T \mathbf{X}_{j(k_j)}\right)^{-1} \mathbf{X}_{j(k_j)}^T \tilde{X}_j$.

(Recall) $\quad \tilde{X}_j \mid \mathbf{X}_{j(k_j)}, a_j^{(k_j)}, d_j, k_j \overset{ind.}{\sim} N_n\left(\mathbf{X}_{j(k_j)} a_j^{(k_j)},\ d_j I_n\right).$

LANCE (LocAl depeNdence CholEsky) prior

$$\pi(A_n, D_n) \quad = \quad \pi(d_1) \prod_{j=2}^{p} \pi(a_j^{(k_j)} \mid d_j, k_j)\, \pi_j(k_j)\, \pi(d_j)$$

(iii) Impose Jeffreys' priors for d_j's:

$$\pi(d_j) \quad \overset{i.i.d.}{\propto} \quad d_j^{-1}.$$

α-fractional posterior

- α-fractional posterior with power $\alpha \in (0, 1)$:

$$\pi_\alpha(A_n, D_n \mid \mathbf{X}_n) \quad \propto \quad L_n(A_n, D_n)^\alpha \pi(A_n, D_n).$$

- (Advantages)
  - It is robust to model misspecification (Grünwald and van Ommen, 2017).
  - It has nice theoretical properties under relatively weaker conditions compared to the actual posterior (Martin et al., 2017; Bhattacharya et al., 2018).

α-fractional posterior

The induced α-fractional posterior has a closed form:

$$
\begin{aligned}
a_j^{(k_j)} \mid d_j, k_j, \mathbf{X}_n \ &\overset{ind}{\sim} \ N_{k_j}\left(\widehat{a}_j^{(k_j)}, \ \frac{d_j}{(\alpha+\gamma)}\left(\mathbf{X}_{j(k_j)}^T \mathbf{X}_{j(k_j)}\right)^{-1}\right), \quad j = 2, \ldots, p, \\
d_j \mid k_j, \mathbf{X}_n \ &\overset{ind}{\sim} \ IG\left(\frac{\alpha n}{2}, \ \frac{\alpha n}{2}\widehat{d}_j^{(k_j)}\right), \quad j = 1, \ldots, p, \\
\pi_\alpha(k_j \mid \mathbf{X}_n) \ &\propto \ p^{-c_1 k_j}\left(1 + \frac{\alpha}{\gamma}\right)^{-\frac{k_j}{2}} \left(\widehat{d}_j^{(k_j)}\right)^{-\frac{\alpha n}{2}}, \quad j = 2, \ldots, p,
\end{aligned}
$$

where $\widehat{d}_j^{(k_j)} = n^{-1}\tilde{X}_j^T(I_n - H_{\mathbf{X}_{j(k_j)}})\tilde{X}_j$ and $H_{\mathbf{X}_{j(k_j)}} = \mathbf{X}_{j(k_j)}(\mathbf{X}_{j(k_j)}^T \mathbf{X}_{j(k_j)})^{-1}\mathbf{X}_{j(k_j)}^T$.

We use $\alpha = 0.99$, $\gamma = 0.1$ and choose c_1 based on Bayesian cross-validation.

Main results

Main results 1

Theorem (Bandwidth selection consistency)

Let Ω_{0n} *and* k_{0j} *be the true precision matrix and bandwidth for the* j*th row.*

Assume that the true Cholesky factor Ω_{0n} *satisfies the regularity conditions.*

If $k_0 \log p \le cn$*, where* $k_0 = \max_{2 \le j \le p} k_{0j}$ *for some constant* $c > 0$*, then we have*

$$\mathbb{E}_0\Big\{\pi_\alpha\Big(k_j = k_{0j} \ \text{ for all } \ 2 \le j \le p \mid \mathbf{X}_n\Big)\Big\} \quad \longrightarrow \quad 1 \quad \text{ as } n \to \infty.$$

Main results 2

Theorem (Posterior convergence rate)

Let A_{0n} be the true Cholesky factor. Suppose that the conditions in the above theorem hold. If $k_0 \log p = o(n)$, then we have

$$
\begin{aligned}
\mathbb{E}_0\Big[\pi_\alpha\Big\{\|A_n - A_{0n}\|_{\max} \geq K_{\mathrm{chol}}\Big(\frac{k_0 + \log p}{n}\Big)^{\frac{1}{2}} \mid \mathbf{X}_n\Big\}\Big] &= o(1),\\
\mathbb{E}_0\Big[\pi_\alpha\Big\{\|A_n - A_{0n}\|_{\infty} \geq K_{\mathrm{chol}}\sqrt{k_0}\Big(\frac{k_0 + \log p}{n}\Big)^{\frac{1}{2}} \mid \mathbf{X}_n\Big\}\Big] &= o(1),\\
\mathbb{E}_0\Big[\pi_\alpha\Big\{\|A_n - A_{0n}\|_F^2 \geq K_{\mathrm{chol}}\frac{\sum_{j=2}^{p}(k_{0j} + \log j)}{n} \mid \mathbf{X}_n\Big\}\Big] &= o(1)
\end{aligned}
$$

as $n \to \infty$, for some constant $K_{\mathrm{chol}} > 0$.

Simulation

Bayesian cross-validation

- We propose to choose the hyperparameter c_1 based on Bayesian cross-validation (Gelman et al. 2014).
- For a given hyperparameter c_1, the estimated out-of-sample log predictive density is

$$\begin{aligned}\mathrm{lpd}_{\mathrm{cv}}(c_1) &= \sum_{\nu=1}^{n_{\mathrm{cv}}} \log f_{c_1}(\mathbf{X}_{I_2(\nu)} \mid \mathbf{X}_{I_1(\nu)}) \\ &= \sum_{\nu=1}^{n_{\mathrm{cv}}} \log \Big\{ \sum_k f(\mathbf{X}_{I_2(\nu)} \mid k) \pi_{\alpha,c_1}(k \mid \mathbf{X}_{I_1(\nu)}) \Big\},\end{aligned}$$

  where $k = (k_2, \ldots, k_p)$.
- The aim of the Bayesian cross-validation is to find the optimal c_1 maximizing $\mathrm{lpd}_{\mathrm{cv}}(c_1)$:

$$\hat{c}_1 = \underset{c_1}{\operatorname{argmax}}\ \mathrm{lpd}_{\mathrm{cv}}(c_1).$$

Bayesian cross-validation

- Note that

$$\begin{aligned}
\pi_{\alpha,c_1}(k_j \mid \mathbf{X}_{I_1(\nu)}) &\propto p^{-c_1 k_j}\left(1+\frac{\alpha}{\gamma}\right)^{-\frac{k_j}{2}}\left\{\widehat{d}_j^{(k_j)}(I_1(\nu))\right\}^{-\frac{\alpha n_1}{2}}, \\
f(\mathbf{X}_{I_2(\nu)} \mid k) &= \prod_{j=2}^{p}\left[\pi^{-\frac{n_2}{2}}\Gamma\Big(\frac{n_2}{2}\Big)\Big(1+\frac{1}{\gamma}\Big)^{-\frac{k_j}{2}}\left\{\widehat{d}_j^{(k_j)}(I_2(\nu))\right\}^{-\frac{n_2}{2}}\right],
\end{aligned}$$

  where $\widehat{d}_j^{(k_j)}(I_1(\nu))$ is the estimated variance $\widehat{d}_j^{(k_j)}$ using $\mathbf{X}_{I_1(\nu)}$.
- When calculating $\mathrm{lpd}_{\mathrm{cv}}(c_1)$, the main computational burden comes from calculating $\widehat{d}_j^{(k_j)}(I_1(\nu))$ and $\widehat{d}_j^{(k_j)}(I_2(\nu))$ for each k_j and j.
- Therefore, LANCE prior enables scalable cross-validation-based inference even in high-dimensions.

Simulation settings

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} N_p(0, \Omega_{0n}^{-1})$ with $\Omega_{0n} = (I_p - A_{0n})^T D_{0n}^{-1} (I_p - A_{0n})$.

- Each nonzero elements in $A_{0n} = (a_{0,jl})$ is drawn independently from

  $$a_{0,jl} = S_{jl} Z_{jl},$$

  where $\mathbb{P}(S_{jl} = -1) = \mathbb{P}(S_{jl} = 1) = 0.5$ and $Z_{jl} \overset{i.i.d.}{\sim} Unif\big([A_{0,\min}, A_{0,\max}]\big)$.

  The remaining entries were set to zero.

- $D_{0n} = diag(d_{0j})$, where $d_{0j} \overset{i.i.d.}{\sim} Unif[2, 5]$.

Simulation settings

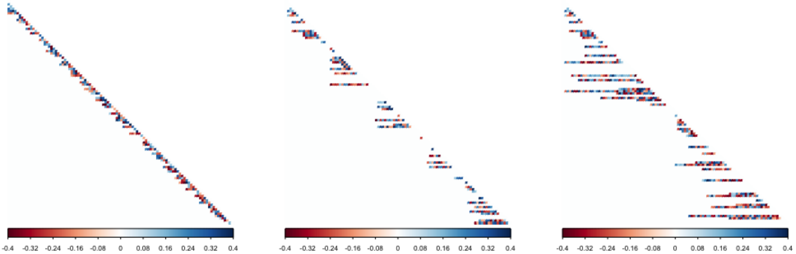


Figure: The true Cholesky factors for Model 1 (Left), Model 2 (Middle) and Model 3 (Right).

Simulation results

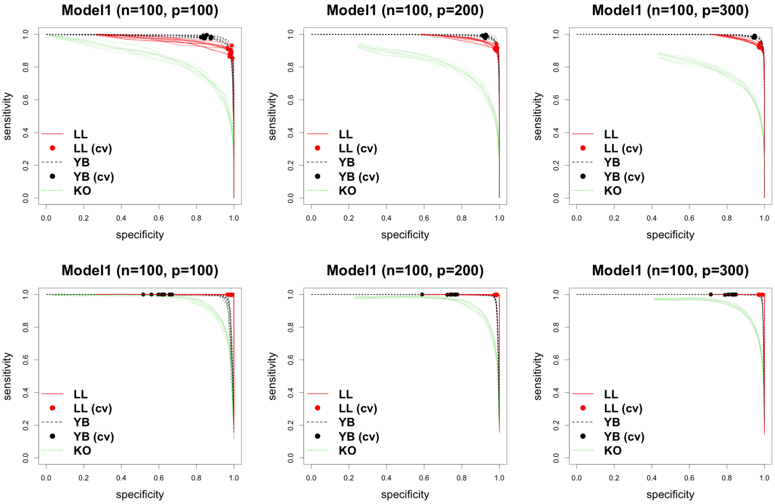


Figure: ROC curves based on 10 simulated data sets from Model 1.

Top row: $(A_{0,\min}, A_{0,\max}) = (0.1, 0.4)$ / Bottom row: $(A_{0,\min}, A_{0,\max}) = (0.4, 0.6)$

Simulation results

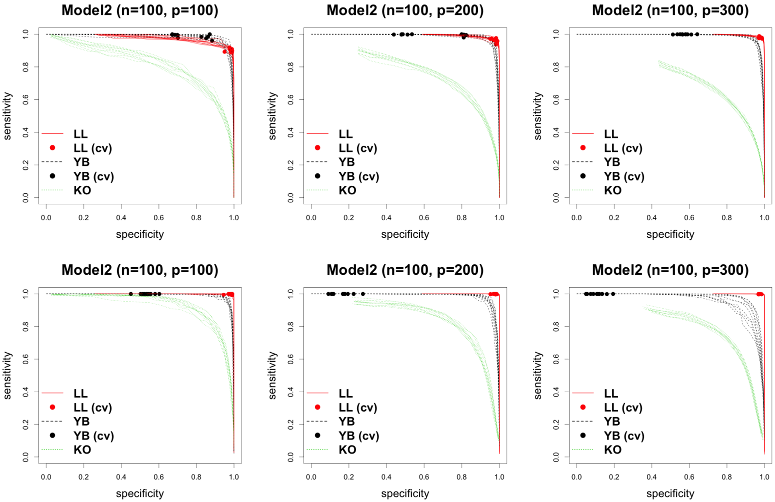


Figure: ROC curves based on 10 simulated data sets from Model 2.

Top row: $(A_{0,\min}, A_{0,\max}) = (0.1, 0.4)$ / Bottom row: $(A_{0,\min}, A_{0,\max}) = (0.4, 0.6)$

Simulation results

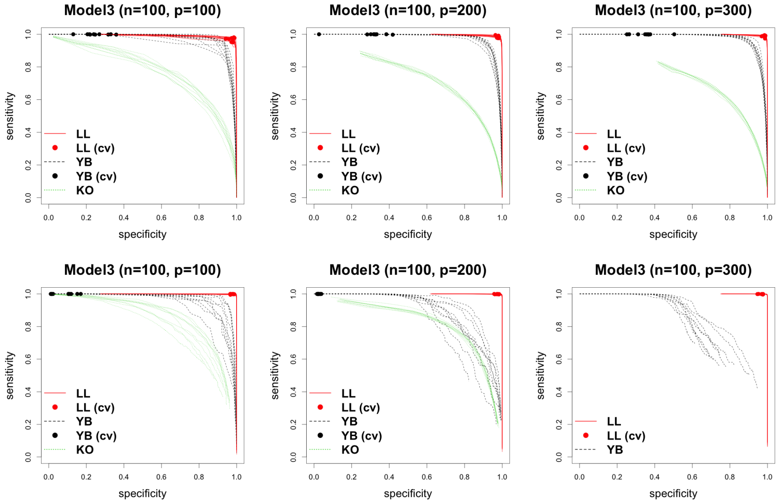


Figure: ROC curves based on 10 simulated data sets from Model 3.

Top row: $(A_{0,\min}, A_{0,\max}) = (0.1, 0.4)$ / Bottom row: $(A_{0,\min}, A_{0,\max}) = (0.4, 0.6)$

Simulation results

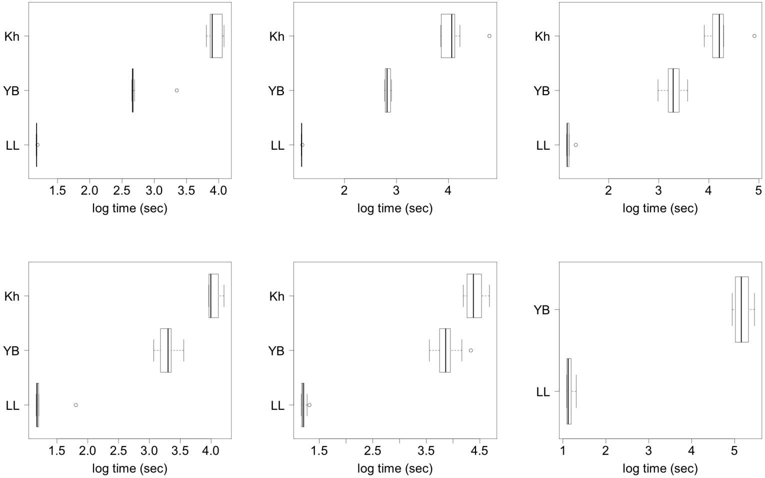


Figure: Log computation times for each setting with $n = 100$ and $p = 300$.

Summary

Summary

- We proposed a new Bayesian procedure for high-dimensional local dependence learning.
- The induced posterior allows a fast computation, which enables scalable inference for large data set.
- Our theoretical result loosens the required conditions on dimensionality, sparsity, the beta-min condition for the Cholesky factors.
- Future work
  - DAG models with unknown ordering
  - Time-varying dependence structure

References I

Baiguo An, Jianhua Guo, and Yufeng Liu, *Hypothesis testing for band size detection of high-dimensional banded precision matrices*, Biometrika **101** (2014), no. 2, 477–483.

Sayantan Banerjee and Subhashis Ghosal, *Posterior convergence rates for estimating large precision matrices using graphical models*, Electronic Journal of Statistics **8** (2014), no. 2, 2111–2137.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang, *Bayesian fractional posteriors*, The Annals of Statistics (2018), to appear.

Xuan Cao, Kshitij Khare, and Malay Ghosh, *Posterior graph selection and estimation consistency for high-dimensional bayesian dag models*, The Annals of Statistics **47** (2019), no. 1, 319–348.

Andrew Gelman, Jessica Hwang, and Aki Vehtari, *Understanding predictive information criteria for bayesian models*, Statistics and computing **24** (2014), no. 6, 997–1016.

References II

Peter Grünwald, Thijs van Ommen, et al., *Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it*, Bayesian Analysis **12** (2017), no. 4, 1069–1103.

Kshitij Khare, Sang-Yun Oh, Syed Rahman, and Bala Rajaratnam, *A scalable sparse cholesky based approach for learning high-dimensional covariance matrices in ordered data*, Machine Learning **108** (2019), no. 12, 2061–2086.

Kyoungjae Lee and Jaeyong Lee, *Estimating large precision matrices via modified cholesky decomposition*, Statistica Sinica (2017), no. accepted.

Kyoungjae Lee and Lizhen Lin, *Bayesian bandwidth test and selection for high-dimensional banded precision matrices*, Bayesian Analysis **15** (2020), no. 3, 737–758.

Kyoungjae Lee, Jaeyong Lee, and Lizhen Lin, *Minimax posterior convergence rates and model selection consistency in high-dimensional dag models based on sparse cholesky factors*, The Annals of Statistics **47** (2019), no. 6, 3413–3437.

References III

Ryan Martin, Raymond Mess, Stephen G Walker, et al., *Empirical Bayes posterior concentration in sparse high-dimensional linear models*, Bernoulli **23** (2017), no. 3, 1822–1847.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and Ji Zhu, *Sparse permutation invariant covariance estimation*, Electronic Journal of Statistics **2** (2008), 494–515.

Sara van de Geer and Peter Bühlmann, ℓ_0*-penalized maximum likelihood for sparse directed acyclic graphs*, The Annals of Statistics **41** (2013), no. 2, 536–567.

Guo Yu and Jacob Bien, *Learning local dependence in ordered data*, Journal of Machine Learning Research **18** (2017), no. 42, 1–60.