Post-Processed Posterior for Banded Covariances

Jaeyong Lee

Department of Statistics
Seoul National University

June 12th 2019

Jointly with
Kwangmin Lee (Ph.D. student in Seoul National University) and
Kyoungjae Lee (Inha University, South Korea).

Content

Content

Unconstrained Covariance Inference

- (Model)
$$X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N_p(0, \Sigma).$$
- (Goal) Inference on Σ
- (Yang and Berger 1996) $IW_p(0_{p\times p}, p+1)$ is the Jeffreys prior, where $0_{p\times p}$ is the $p \times p$-dimensional zero matrix.
- (Huang et al. 2013) If $\Sigma = (\sigma_{ij}) \sim IW_p(\text{diagonal matrix}, 2p+1)$,
$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \sim Unifom(-1, 1).$$
- (Lee and Lee 2018)
  - When $p \leq cn$, $0 \leq c < 1$, the inverse-Wishart prior has the minimax optimal rate under the spectral norm.
  - When $p \geq n$, $\delta_{I_p}$is the prior with minimax optimal rate. Thus, in this case, the inference without restriction is hopeless. Certain restriction is necessary for the inference of Σ.

Banded Covariance Inference

Bayesian side

- (Khare et al., 2011) Covariance graphical model can be used for the inference of the banded covariance.
- (Pros and cons)
  - It provides interval estimators for any functional of covariance.
  - Its computation can be slow and sometimes unstable.
  - It lacks an asymptotic justification.

Frequentist side

- (Bickel and Levina, 2008) The banding estimator is proposed and its convergence rate is obtained.
- (Wu and Pourahmadi, 2009) The banding estimator for stationary time series.
- (Wiesel and Globerson 2012) The banding estimator for time-varying autoregressive moving average model.
- (Pros and cons)
  - The computation of banding estimator is fast.
  - It lacks interval estimators.
  - No minimax theory for banded covariances. There is an minimax results for bandable covariances (Cai and Zhou 2010, 2012).

Banded Covariance Inference

Bayesian side

- (Khare et al., 2011) Covariance graphical model can be used for the inference of the banded covariance.
- (Pros and cons)
  - It provides interval estimators for any functional of covariance.
  - Its computation can be slow and sometimes unstable.
  - It lacks an asymptotic justification.

Goal

We want a method for banded covariances that

1. is computationally efficient,
2. provides interval estimators, and
3. has asymptotic optimality.

Frequentist side

- (Bickel and Levina, 2008) The banding estimator is proposed and its convergence rate is obtained.
- (Wu and Pourahmadi, 2009) The banding estimator for stationary time series.
- (Wiesel and Globerson 2012) The banding estimator for time-varying autoregressive moving average model.
- (Pros and cons)
  - The computation of banding estimator is fast.
  - It lacks interval estimator.
  - No minimax theory for banded covariances. There is an minimax results for bandable covariances (Cai and Zhou 2010, 2012).

Post-Processed Posterior

Model

$$X_1, X_2, \ldots X_n | \Sigma \overset{i.i.d.}{\sim} N_p(0, \Sigma),$$

where $\Sigma \in \mathcal{B}_{p,k} = \{\Sigma = (\sigma_{ij}) \in \mathcal{C}_p : \sigma_{ij} = 0 \text{ if } |i-j| > k\}$ and $\mathcal{C}_p$ is the set of all positive definite $p \times p$ matrices.

Initial Prior and Posterior

- Initial prior: $\pi^i(\Sigma) = [\Sigma]_i = IW_p(B_0, \nu_0)$.
- Initial posterior: $\pi^i(\Sigma|\mathbb{X}_n) = [\Sigma|\mathbb{X}_n]_i = IW_p(B_0 + nS_n, \nu_0 + n)$, where $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$.

Post-Processed Posterior

- Post-processing function: $f : \mathcal{C}_p \longrightarrow \mathcal{B}_{p,k}$

$$\begin{aligned} f(\Sigma) &= B_k^{(\epsilon n)}(\Sigma) \\ &= \begin{cases} B_k(\Sigma) + (\epsilon_n - \lambda_{min}(B_k(\Sigma)))I_p, & \lambda_{min}(B_k(\Sigma^{(i)})) < 0 \\ B_k(\Sigma) + \epsilon_n I_p, & \lambda_{min}(B_k(\Sigma)) \geq 0. \end{cases} \end{aligned}$$

- post-processed posterior:

$$\pi^{pp}(\Sigma|\mathbb{X}_n) = [\Sigma|\mathbb{X}]_{pp} := [f(\Sigma)|\mathbb{X}_n]_i.$$

Post-Processing Algorithm

Step 1. (initial posterior computing)

1. Using the initial prior $\pi^i = IW(B_0, \nu_0)$, obtain the initial posterior.

$$\Sigma|\mathbb{X}_n \sim \pi^i(\Sigma|\mathbb{X}_n) = IW(B_0 + nS_n, \nu_0 + n).$$

2. Draw the intial posterior samples:

$$\Sigma^{(1)}, \Sigma^{(2)}, \ldots, \Sigma^{(N)} \overset{i.i.d.}{\sim} \pi^i(\Sigma|\mathbb{X}_n).$$

Step 2. (post-processing)
Draw samples from the post-processed posterior $\pi^{pp}(\Sigma|\mathbb{X}_n; f)$ by transforming the initial posterior samples using the post-processing function f:

$$\begin{aligned} \Sigma_{(i)} &:= f(\Sigma^{(i)}) = B_k^{(\epsilon_n)}(\Sigma^{(i)}) \\ &= \begin{cases} B_k(\Sigma^{(i)}) + (\epsilon_n - \lambda_{min}(B_k(\Sigma^{(i)})))I_p, & \lambda_{min}(B_k(\Sigma^{(i)})) < 0 \\ B_k(\Sigma^{(i)}) + \epsilon_n I_p, & \lambda_{min}(B_k(\Sigma^{(i)})) \geq 0. \end{cases} \end{aligned}$$

Existing Ideas of Transforming Posterior Samples

- (Projected posterior) Projecting posterior samples onto the constraint spaces.
  - the ordered constraint parameter space in generalized linear models (Dunson and Leelon 2003).
  - monotone or unimodal parameter space (Gunn and Dunson 2005).
  - monotone regression using Gaussian process prior (Lin and Dunson 2014).
  - existence of data-dependent prior for the projected posterior (Patra and Dunson 2018)
  - credible sets of piecewise-constant regression with monotone contraint (Chakraborty and Ghosal 2019, BNP 2019 Talk)
- support recovery for sparse precision matrix from the posterior samples of dense model. (Bashir et al. 2018)
- Estimating number of clusters with Dirichlet mixture posterior. (Nguyen, BNP 2019).

- Post-processed posterior for the banded covariance is 'almost' the projected posterior under the Frobenius norm without positive definite condition.
- The idea of of post-processed posterior is the same as the projected posterior. The difference lies in the choice of transform. The post-processing function is chosen with minimax consideration, while the transformation of the projected posterior is the projection onto the constrained space.

An Example of PPP

- $\theta = (\theta_1, \theta_2) \in \mathbb{R}^p$. True value of θ is $\theta_0 = (\theta_{01}, \theta_{02})$.
- $X_1, \ldots, X_n | \theta \overset{i.i.d.}{\sim} p(x|\theta)$. $\mathbb{X}_n = (X_1, \ldots, X_n)$.
- (True model)
  - $\theta_1 \sim \pi(\theta_1)$.
  - $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p(x_i|\theta_1, \theta_2 = 0)$
  - $[\theta_1|\mathbb{X}_n]_T$: posterior under the true model.
- (Post-processed posterior)
  - $\pi^i(\theta_1, \theta_2)$ is the initial prior.
  - $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p(x_i|\theta_1, \theta_2)$
  - $[\theta_1, \theta_2|\mathbb{X}_n]_i$: initial posterior
  - The post-processing function is $f(\theta_1, \theta_2) = (\theta_1, 0)$.
  - $[\theta_1|\mathbb{X}_n]_{PPP} = [\theta_1|\mathbb{X}_n]_i$.
- (Question) Relation between $[\theta_1|\mathbb{X}_n]_{PPP}$ and $[\theta_1|\mathbb{X}_n]_T$?

An Example of PPP: PPP is conservative

- The Bernstein von-Mises theorem holds for the posterior distribution $[\theta_1, \theta_2|\mathbb{X}_n] = \pi^i(\theta_1, \theta_2|\mathbb{X}_n) \propto \pi^i(\theta_1, \theta_2)p(\mathbb{X}_n|\theta_1, \theta_2)$: as $n \longrightarrow \infty$,

$$[\sqrt{n}((\theta_1, \theta_2)^T - (\hat{\theta}_1^*, \hat{\theta}_2^*)^T)|\mathbb{X}_n] \xrightarrow{d} N(0, \mathcal{I}^{-1}(\theta_{01}, 0)),\ \mathbb{P}_{\theta_{01},0} - a.s., \tag{1}$$

  where $(\hat{\theta}_1^*, \hat{\theta}_2^*)^T$ is the maximum likelihood estimator of $(\theta_1, \theta_2)^T$, and $\mathcal{I}(\theta_1, \theta_2)$ is the Fisher-information matrix.

- The Bernstein von-Mises theorem holds for the posterior distribution $[\theta_1|\mathbb{X}_n]_T = \pi^i(\theta_1|\mathbb{X}_n) \propto \pi^i(\theta_1)p(\mathbb{X}_n|\theta_1, 0)$:

$$[\sqrt{n}(\theta_1 - \hat{\theta}_1)|\mathbb{X}_n]_T \xrightarrow{d} N(0, \mathcal{I}_{11}^{-1}(\theta_{01}, 0)), \mathbb{P}_{\theta_{01}} - a.s., \tag{2}$$

  where $\hat{\theta}_1$ is the maximum likelihood estimator of θ_1 and $\mathcal{I}_{11}(\theta_1, 0)$ is the Fisher-information matrix under the true model.

Theorem. Under the above assumptions, the following hold.

(i) $[\sqrt{n}(f(\theta_1, \theta_2) - \hat{\theta}_1^*)|\mathbb{X}_n] = [\sqrt{n}(\theta_1 - \hat{\theta}_1^*)|\mathbb{X}_n]_{PPP} \xrightarrow{d} N(0, \mathcal{I}_{11\cdot 2}^{-1}(\theta_{01}, 0))$

(ii) $\mathcal{I}_{11\cdot 2}^{-1}(\theta_{01}, 0) - \mathcal{I}_{11}^{-1}(\theta_{01}) \geq 0$, where $A - B \geq 0$ means $A - B$ is non-negative definite.

Content

Posterior convergence rate

$\epsilon_n \longrightarrow 0$ is a posterior convergence rate, if

$$\pi(||\theta - \theta_0|| > M_n \epsilon_n | X_n) \longrightarrow 0, \quad \text{in } \mathbb{P}_{\theta_0}, \text{ as } n \longrightarrow \infty$$

for any $M_n \longrightarrow \infty$. $\pi(\cdot|X_n)$ is the posterior of θ given the observation $X_n = (x_1, \ldots, x_n)$ and θ_0 is the true value of θ.

Posterior convergence rate

$\epsilon_n \longrightarrow 0$ is a posterior convergence rate, if

$$\pi(||\theta - \theta_0|| > M_n \epsilon_n | X_n) \longrightarrow 0, \quad \text{in } \mathbb{P}_{\theta_0}, \text{ as } n \longrightarrow \infty$$

for any $M_n \longrightarrow \infty$. $\pi(\cdot|X_n)$ is the posterior of θ given the observation $X_n = (x_1, \ldots, x_n)$ and θ_0 is the true value of θ.

Unsatisfactory features to me

1. M_n makes the definition unclean. But it is necessary for a parametric model to have rate $\dfrac{1}{\sqrt{n}}$.
2. If the posterior consistency does not hold, the posterior convergence rate is not defined. To define it, the loss function needs to be scaled first. The rescaled loss is may not be a natural one.

A Quote from Ghosal and van der Vaart

"It may be noted that we defined 'a rate of convergence' rather than 'the rate of convergence'. Naturally, the most precise assertion corresponds to the smallest possible value of ϵ_n, but in general, existence of the smallest such sequence is questionable. Further, it is often very hard to show that a rate cannot be improved. Thus our obtained rate will actually be an upper bound for the targeted rate. Generally, we shall be happy with 'a rate' which we think is equal to or close to 'the rate'. Indeed, in many applications, we shall obtain the optimal rate of convergence (in comparison with convergence rate of point estimators), which cannot be further improved."
– Ghosal and van der Vaart. (2017). Fundamentals of Nonparametric Bayesian Inference.

Unsatisfactory features of the Posterior convergence rate

- "The rate" of a posterior is not defined, not to mention the optimal rate.
- Optimality is vaguely understood in connection with the minimax rate of estimators.
- The lower bound is difficult to obtain.
- To define convergence rate, the posterior consistency needs to be satisfied. Loss functions without consistency need to be scaled for consistency.

P-loss, P-risk: A framework for Bayesian minimax rate (Lee and Lee 2018)

Decision theory framework

- (parameter and parameter space) $\theta \in \Theta$
- (observation) $X|\theta \sim f(x|\theta)$
- (action and action space) $a \in \mathcal{A}$
- (loss) $L(\theta_0, a)$
- (decision rule) $\delta(X)$. $\delta : \mathcal{X} \longrightarrow \mathcal{A}$
- (risk) $R(\theta_0, \delta) = \mathbb{E}_\theta L(\theta, \delta(X))$

Proposed framework

- (parameter and parameter space) $\theta \in \Theta$
- (observation) $X|\theta \sim f(x|\theta)$
- (action and action space) An action is a probability measure on Θ, i.e. a posterior is an action.
- (decision rule) $\pi \in \Pi$. A prior with an observation determines an action, the posterior.
- (P-loss) $\mathcal{L}(\theta_0, \pi(\cdot|X)) = \mathbb{E}^\pi(L(\theta, \theta_0)|X)$
- (P-risk) $\mathcal{R}(\theta_0, \pi) = \mathbb{E}_{\theta_0}\mathbb{E}^\pi(L(\theta, \theta_0)|X)$

P-risk minimax rate

- r_n is the P-risk (Bayesian) minimax rate if and only if

$$\inf_{\pi \in \Pi} \sup_{\theta_0 \in \Theta^*} \mathbb{E}_{\theta_0} \mathbb{E}^{\pi}(L(\theta, \theta_0) | X_n) \asymp r_n, \text{ as } n \longrightarrow \infty.$$

  - $a_n \asymp b_n$ means that there exist $0 < c < C$ such that $c < a_n / b_n < C$ for all sufficiently large n.
  - $\Theta^* \subset \Theta$.

- A prior π^* attains the P-risk minimax rate if

$$\sup_{\theta_0 \in \Theta^*} \mathbb{E}_{\theta_0} \mathbb{E}^{\pi^*}(L(\theta, \theta_0) | X_n) \asymp r_n, \text{ as } n \longrightarrow \infty.$$

Advantages of P-risk minimax rate

- The convergence rate of a posterior is defined in a mathematically clean way. Recall "the" convergence rate of a posterior was not defined.
- The minimax rate of varying parameter spaces and varying prior classes can be defined in an obvious fashion.
- Non-traditional posteriors or inference functions can be considered in the P-risk framework.

- P-risk is already being used, e.g. Kim and Lee (2001), Jang, Lee, and Lee (2010), Lee and Oh (2013), Castillo and van der Pas (O'Bayes 2019).

Two simple facts

Theorem
The convergence rate of $\mathcal{R}(\theta_0, \pi)$ is a posterior convergence rate.

Theorem
The frequentist minimax lower bound for θ_0 is also a Bayesian minimax lower bound.

Content

Convergence Rate for Banded Covariances

Theorem. (Convergence rate of the post-processed posterior for banded covariances)
Let the prior π^i of Σ be $IW_p(A_n, \nu_n)$. If $\nu_n - 2p = o(n)$, $A_n \in \mathcal{B}_{p,k}$ and $||A_n||^2 = o(n)$, then for all sufficiently large n,

$$\sup_{\Sigma_0 \in \mathcal{B}_{p,k}} E_{\Sigma_0}\{E^{\pi^i}(||B_k^{(\epsilon_n)}(\Sigma) - \Sigma_0||^2 \mid \mathbb{X}_n)\} \le C(\log k)^2 \frac{k + \log p}{n},$$

where $\epsilon_n^2 = O\{(\log k)^2(k + \log p)/n\}$ in $f = B_k^{(\epsilon_n)}$ and $\mathcal{B}_{p,k}$ has bounded eigenvalue conditions.

Theorem. (Minimax lower bound for banded covariances)
Suppose $p \le \exp(\gamma n)$ for some constant $\gamma > 0$. Then, for all sufficiently large n,

$$\inf_{(\pi,f)\in\Pi^*} \sup_{\Sigma_0 \in \mathcal{B}_{p,k}} E_{\Sigma_0}\{\mathbb{E}^{\pi}(||f(\Sigma) - \Sigma_0||^2 \mid \mathbb{X}_n)\} \ge c \frac{k + \log p}{n}.$$

$\mathcal{B}_{p,k} := \mathcal{B}_{p,k}(M_0, M_1) = \{\Sigma \in \mathcal{C}_p : \sigma_{ij} = 0 \text{ if } |i-j| > k, \forall i, j \in [p], \lambda_{\max}(\Sigma) \le M_0, \lambda_{\min}(\Sigma) \ge M_1\}$, $0 < M_1 < M_0$, $[p] = \{1, 2, \ldots, p\}$, and $\mathcal{C}_p$ is the set of all $p \times p$-dimensional positive definite matrices.

Remarks

$$\sup_{\Sigma_0 \in \mathcal{B}_{p,k}} E_{\Sigma_0} \{ E^{\pi^i} (||B_k^{(\epsilon_n)}(\Sigma) - \Sigma_0||^2 \mid \mathbb{X}_n) \} \leq C (\log k)^2 \frac{k + \log p}{n},$$

$$\inf_{(\pi, f) \in \Pi^*} \sup_{\Sigma_0 \in \mathcal{B}_{p,k}} E_{\Sigma_0} \{ \mathbb{E}^{\pi} (||f(\Sigma) - \Sigma_0||^2 \mid \mathbb{X}_n) \} \geq c \frac{k + \log p}{n}.$$

- The PPP is almost minimax optimal.
  The upper bound of the P-risk is almost the same as the lower bound except $(\log k)^2$ factor.
- The minimax result includes proper Bayesian procedures,
  i.e. proper Bayesian posteriors can not do better than the PPP (except $(\log k)^2$ factor).

Convergence Rate for Bandable Covariances

Theorem. (Convergence rate of the post-processed posterior for bandable covariances)
Let prior π^i of Σ be $IW_p(A_n, \nu_n)$. If $\nu_n - 2p = o(n)$, $A_n \in \mathcal{B}_{p,k}$ and $||B_k(A_n)||^2 = o(n)$, then for all sufficiently large n,

$$\sup_{\Sigma_0 \in \mathcal{F}_\alpha} E_{\Sigma_0}\{E^{\pi^i}(||B_k^{(\epsilon_n)}(\Sigma) - \Sigma_0||^2 \mid \mathbb{X}_n)\} \leq C \min\left\{(\log k)^2 \frac{\log p + k}{n} + k^{-2\alpha}, \frac{p}{n}\right\},$$

where $\epsilon_n^2 = O[\min\{(\log k)^2(k + \log p)/n + k^{-2\alpha},\ p/n\}]$.

Theorem. (Minimax lower bound for bandable covariances)
For all sufficiently large n,

$$\inf_{(\pi,f)\in\Pi^*} \sup_{\Sigma_0 \in \mathcal{F}_\alpha} E_{\Sigma_0} E^{\pi}\left[||f(\Sigma) - \Sigma_0||^2 \mid \mathbb{X}_n\right] \geq C \min\left\{n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n}, \frac{p}{n}\right\}.$$

$$\mathcal{F}_\alpha = \left\{\Sigma = (\sigma_{ij}) \in \mathcal{C}_p : \sum_{(i,j):|i-j|\geq k} |\sigma_{ij}| \leq Mk^{-\alpha}, \forall k \geq 1, \lambda_{max}(\Sigma) \leq M_0, \lambda_{min}(\Sigma) \geq M_1\right\},$$

with $\alpha, M, M_0, M_1 > 0$.

Content

Banded Covariances

A Framework for Bayesian minimax rate

Minimaxity

Simulation Studies

A Simulation Study

- Let $\Sigma_0^{(4)} = (\sigma_{0,ij}^{(4)})_{p\times p}$, where

$$\sigma_{0,ij}^{(4)} = \begin{cases} 1, & 1 \le i = j \le p \\ \rho|i-j|^{-(\alpha+1)}, & 1 \le i \ne j \le p, \end{cases}$$

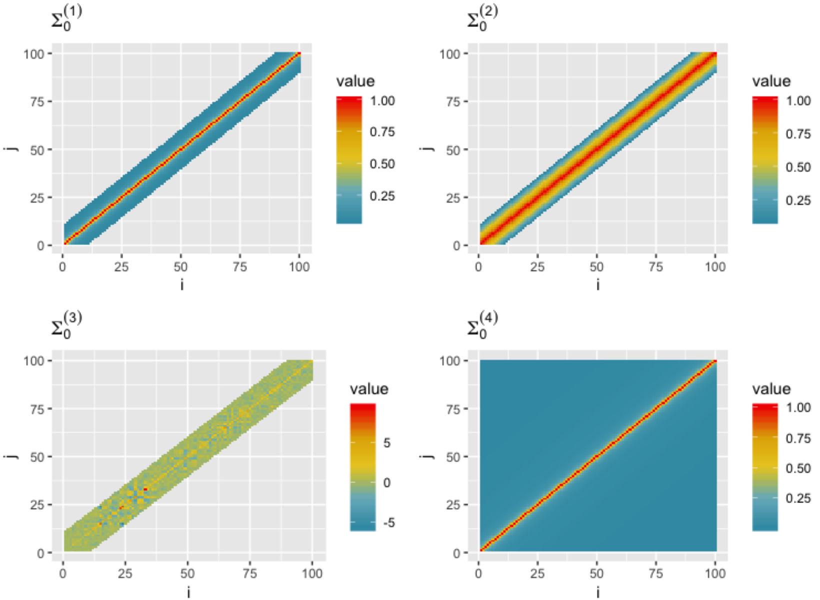
  $\rho = 0.6$ and $\alpha = 0.1$.

- $\Sigma_0^{(1)} = B_{k_0}(\Sigma_0^{(4)}) + [\epsilon - \{\lambda_{min}(B_{k_0}(\Sigma_0^{(4)})) \wedge 0\}]I_p$, where k_0 is the bandwidth and $\epsilon = 10^{-4}$.
- $\Sigma_0^{(2)} = (\sigma_{0,ij}^{(2)})_{p\times p}$, where $\sigma_{0,ij}^{(2)} = \{1 - |i-j|/(k_0+1)\} \wedge 0, 1 \le i,j \le p$.
- $\Sigma_0^{(3)} = L_0 D_0 L_0^T$, where

$$L_{ij}^0 = \begin{cases} 1, & 1 \le i = j \le p \\ l_{ij}, & 0 < i - j \le k_0 \\ 0, & \text{otherwise}, \end{cases}$$

  l_{ij} are independent sample from $N(0,1)$, and $D_0 = diag(d_{ii})$ is diagonal matrix where d_{ii} is independent sample from $IG(5,1)$, the inverse-gamma distribution with the shape parameter 5 and the scale parameter 1.

A Simulation Study



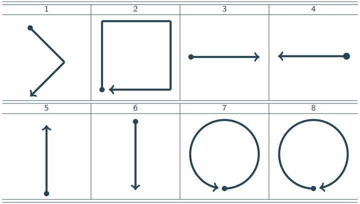
A Simulation Study

Errors of point estimators for banded covariances $\Sigma_0^{(1)}$, $\Sigma_0^{(2)}$ and $\Sigma_0^{(3)}$. Error is defined as the spectral norm of the difference between the true value and the point estimator. For 1–3rd rows, each represents the error by its method. The 4th row represents the error by Khare et al.'s method without outlier, which is defined as an observation that has value over 50. The 5th row represents the number of the outliers.

	$n=50, p=50, k=10$			$n=50, p=100, k=10$		
	$\Sigma_0^{(1)}$	$\Sigma_0^{(2)}$	$\Sigma_0^{(3)}$	$\Sigma_0^{(1)}$	$\Sigma_0^{(2)}$	$\Sigma_0^{(3)}$
Bickel and Levina's method	2.02	3.23	5.77	2.46	4.09	8.66
Post-processed posterior	1.67	3.17	4.39	1.89	3.56	6.95
Khare et al.'s method	1.70	1.484×10^3	7.77×10^5	1.86	3.66×10^8	7.12×10^7
Khare et al.'s method without outliers	1.70	3.11	3.88	1.86	3.34	6.56
(Number of outliers in Khare et al.'s method)	(0)	(6)	(95)	(0)	(84)	(33)

Recognition of Gestures

Data



- An observation corresponds to one of eight motions.
- An observation is a 945-dimensional vector containing accelerations in x, y, z axes at 315 time points.

Analysis

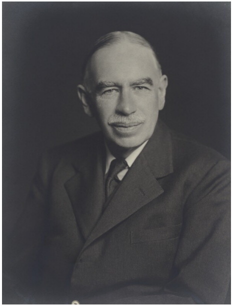
- Linear discrminant analysis with fixed means for each group.
- Covariance is estimated by the banding estimator and the mean of post-processing posterior.

Performance of Linear discriminant analysis

Method	Accuracy
Sample covariance	56.03%
inverse-Wishart posterior	60.13%
Bickel and Levina's method	84.37%
Post-processed posterior	85.73%

When the facts change, I change my opinion. What do you do, Sir?

- John Maynard Keynes



Thank you.