Covariance estimation using generalized fiducial inference

Wen Jenny Shi
joint work with Jan Hannig

Department of Statistics and Operations Research, UNC-Chapel Hill

August 5, 2013

Motivation

- Estimation of Σ is a fundamental problem.
- Distribution for Σ is appealing.
- Fixed zeros in A is a different type of sparsity.
- Sparse A makes sense in some biological/genetic context.
- Choosing priors can be infeasible when the sparse structure of A is unknown.

Generalized fiducial inference (Hannig et al 2006, 2009)

- *Data generating equation*: $X = G(U, \theta)$

Generalized fiducial inference (Hannig et al 2006, 2009)

- *Data generating equation*: $X = G(U, \theta)$
- Set-value function: $Q(x, u) = \{\theta : x = G(u, \theta)\}$

Generalized fiducial inference (Hannig et al 2006, 2009)

- *Data generating equation*: $X = G(U, \theta)$
- Set-value function: $Q(x, u) = \{\theta : x = G(u, \theta)\}$
- *Generalized Fiducial Distribution (GFD)*: conditional distribution of

$$\lim_{\epsilon \to 0} [Q_{x,\epsilon}(U^\star) | \{Q_{x,\epsilon}(U^\star) \neq \emptyset\}]$$

where $Q_{x,\epsilon}(U^\star) = \{\theta : \|x - G(U^\star, \theta)\|_{L_\infty} < \epsilon\}$.

Generalized fiducial inference (Hannig et al 2006, 2009)

- *Data generating equation*: $X = G(U, \theta)$
- Set-value function: $Q(x, u) = \{\theta : x = G(u, \theta)\}$
- *Generalized Fiducial Distribution (GFD)*: conditional distribution of

$$\lim_{\epsilon \to 0} [Q_{x,\epsilon}(U^\star) | \{Q_{x,\epsilon}(U^\star) \neq \emptyset\}]$$

where $Q_{x,\epsilon}(U^\star) = \{\theta : \|x - G(U^\star, \theta)\|_{L_\infty} < \epsilon\}$.

- $r(\theta) = \frac{f(x,\theta)J(x,\theta)}{\int_\Theta f(x,\theta')J(x,\theta')d\theta'}$, where
$J(x, \theta) = \sum_{\substack{\mathbf{i}=(i_1,\cdots,i_p) \\ 1 \le i_1 < \cdots < i_p \le p}} \left| \det \left[\left(\frac{d}{d\theta} G(u, \theta) \Big|_{u=G^{-1}(x,\theta)} \right)_{\mathbf{i}} \right] \right|$.

GFD for covariate A

Data generating function:

$$Y_i = AZ_i, \ i = 1, \cdots, n,$$

where $Z_i \overset{\text{iid}}{\sim} N(0, I)$.
$\Rightarrow$ GFD of A:

$$r(A) \propto J(\mathbf{Y}, A) f(\mathbf{Y}, A),$$

where

$$f(\mathbf{Y}, A) = (2\pi)^{-\frac{np}{2}} |\det(A)|^{-n} \exp\left[-\frac{1}{2}\mathrm{tr}\{nS_n(AA^T)^{-1}\}\right],$$

S_n is the sample covariance matrix.

GFD for covariate A

Data generating function:

$$Y_i = AZ_i, \; i = 1, \cdots, n,$$

where $Z_i \overset{\text{iid}}{\sim} N(0, I)$.

$\Rightarrow$ GFD of A:

$$r(A) \propto J(\mathbf{Y}, A) f(\mathbf{Y}, A),$$

where

$$f(\mathbf{Y}, A) = (2\pi)^{-\frac{np}{2}} |\det(A)|^{-n} \exp\left[-\frac{1}{2}\mathrm{tr}\{nS_n(AA^T)^{-1}\}\right],$$

S_n is the sample covariance matrix.

What about $J(\mathbf{Y}, A)$?

Simplified Jacobian

$$J(\mathbf{Y}, A) = \prod_{i=1}^{p} \binom{p}{p_i} \cdot \prod_{i=1}^{p} \overline{\left|\det\left[(V_i)_{\mathbf{i}_i}\right]\right|},$$

where

- p_i =number of non fixed zeros in row i;
- V_i = a submatrix of V, uniquely determined by the fixed zeros in the i^{th} row of A;
- $V = (Y_1, \cdots, Y_n)^T$;
- $\mathbf{i}_i = (i_1, \cdots, i_{p_i p}),\ s.t.\ 1 \le i_1 < \cdots < i_{p_i p} \le np$;
- $\overline{\left|\det\left[(V_i)_{\mathbf{i}_i}\right]\right|}$ = average absolute determinant of all possible expressions of $(V_i)_{\mathbf{i}_i}$ for a fixed i.

GFD of Σ?

- If there is no fixed zero in A,

$$J(\mathbf{Y}, A) = \left(\sum_{\substack{\mathbf{i}=(i_1,\cdots,i_p) \\ 1\le i_1<\cdots<i_p\le p}} \left|\det\left[(V)_{\mathbf{i}}\right]\right| \right)^p \left|\det(A)\right|^{-p}$$

$$\Rightarrow \Sigma = AA^T \sim \text{Inverse - Wishart } (nS_n, n).$$

Covariance Σ can be estimated using standard Markov chain Monte Carlo (MCMC) methods.

GFD of Σ?

- If there is no fixed zero in A,

$$J(\mathbf{Y}, A) = \left(\sum_{\substack{\mathbf{i}=(i_1,\cdots,i_p) \\ 1 \le i_1 < \cdots < i_p \le p}} \left| \det \left[(V)_{\mathbf{i}} \right] \right| \right)^p \left| \det(A) \right|^{-p}$$

$$\Rightarrow \Sigma = AA^T \sim \text{Inverse - Wishart } (nS_n, n).$$

  Covariance Σ can be estimated using standard Markov chain Monte Carlo (MCMC) methods.

- If there are fixed zeros in A, the fiducial distribution of Σ is usually complicated!

Estimate A and infer $\Sigma = AA^T$

Recall

$$r(A) \propto J(\mathbf{Y}, A) f(\mathbf{Y}, A),$$

$$\Sigma = AA^T.$$

Covariate A can be estimated using

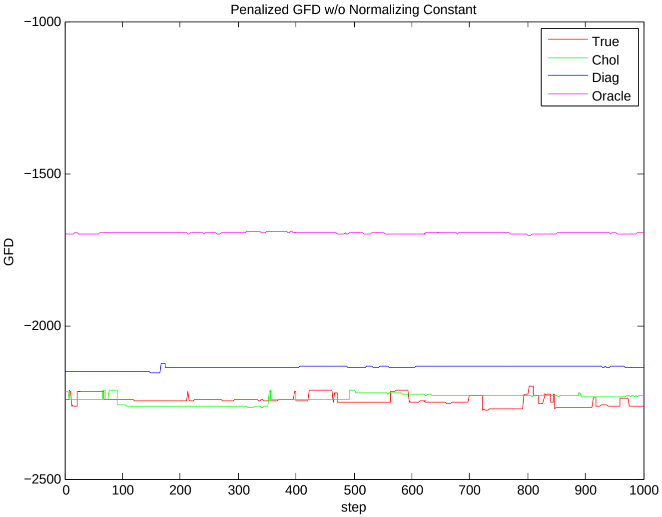
- reversible jump MCMC with moves: *birth, death, update,*

$+$

- Minimum Description Length (MDL) as penalty:

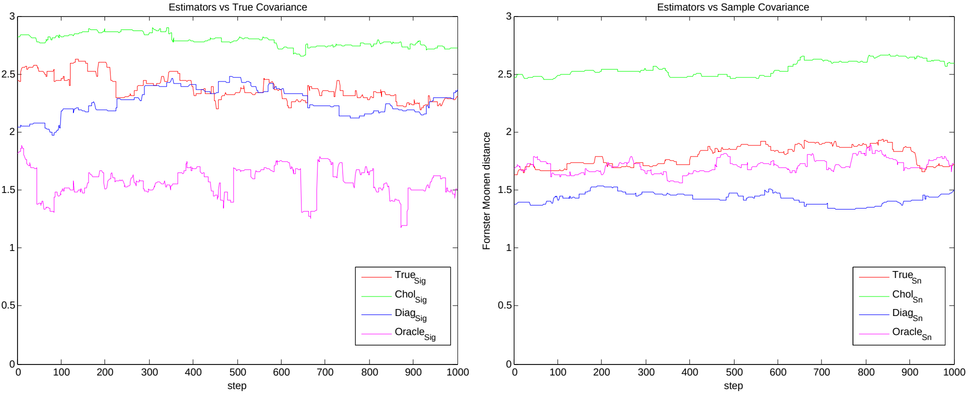
$$p(A) = \sum_{i=1}^{n} \frac{p_i^2}{2} \log n$$

Penalized GFD trace plot



Forstner & Moonen (1999):

$$\mathbf{d}(M, N) = \sqrt{\sum_{i=1}^{n} \ln^2 \lambda_i(M, N)}.$$



Conclusion and future work

- GFI approach to covariance estimation is feasible.
- Implement geometric structure on A.
- Improve the efficiency of the reversible jump MCMC's.

The End.

Thank you!

Questions/suggestions?