Flexible Wishart distributions and their applications

Hélène Massam

York University

with co-authors, G. Letac, B. Rajaratnam, C. Carvalho

Outline

- The Gaussian model
- The graphical Gaussian model $\mathcal{N}_G$
- The W_{Q_G} and W_{P_G} Wisharts
- The hyper Markov property
- The expected values of W_{Q_G} and IW_{P_G}
- Decision theoretic estimation of Σ
- Another Wishart?

Inference for the saturated Gaussian model

$$\mathcal{N} = \{N_r(0,\Sigma),\ \ \Sigma \in P^+\}.$$

where P^+ is the cone of positive definite matrices.
Let $Z_1,\ldots,Z_n$ be a sample from $N_r(0,\Sigma)$.

- The MLE of $n\Sigma$ is $n\tilde{\Sigma} = U = \sum_{i=1}^n Z_i Z_i^t$ with

$$U \sim W_r(\frac{n}{2},\Sigma)$$

- The conjugate prior on $\Omega = \Sigma^{-1}$ is the Wishart

$$\Omega \sim W_r(\alpha,\theta)$$

The Wishart distribution

Recall $\mathcal{F}_\alpha = \{W_r(\alpha, \Sigma),\ \Sigma \in P^+\}$ is the NEF generated by

$$\mu_\alpha(dx) = (\det x)^{\alpha - \frac{r+1}{2}} \mathbf{1}_{P^+}(dx), \text{ with } \alpha > \frac{r-1}{2}$$

and the density of the $W_r(\alpha, \Sigma)$ is

$$W_r(\alpha, \Sigma)(dx) = \frac{(\det \Sigma)^{-\alpha}}{\Gamma_r(\alpha)} \exp -\frac{1}{2}\langle x, \Sigma^{-1}\rangle \ \mu_\alpha(dx)$$

while the density of $Y = X^{-1}$ is the inverse Wishart

$$IW_r(\alpha, \Sigma; dy) = \frac{(\det \Sigma)^{-\alpha}}{\Gamma_r(\alpha)} \exp -\frac{1}{2}\langle y^{-1}, \Sigma^{-1}\rangle \ (\det y)^{-\alpha - \frac{r+1}{2}} \mathbf{1}_{P^+}(dy)$$

Note: Only one shape parameter α

Outline

- The Gaussian model
- The graphical Gaussian model
- The W_{Q_G} and W_{P_G} Wisharts
- The hyper Markov property
- The expected values of W_{Q_G} and IW_{P_G}
- Decision theoretic estimation of Σ
- Another Wishart?

Graphical models: decomposable graphs

Let $G = (V, E)$ be a graph: V is the set of vertices and E the set of edges.

G is said to be decomposable if it does not contain a cycle of length greater than or equal to 4 without a chord.

Example $\overset{1}{\bullet} - \overset{2}{\bullet} - \overset{3}{\bullet} - \overset{4}{\bullet}$

$C_1 = \{1, 2\}, C_2 = \{2, 3\}, C_3 = \{3, 4\}$ are the cliques

$S_2 = \{2\}, S_3 = \{3\}$ are the separators.

Perfect ordering of the cliques

A graph is decomposable if and only if there is a perfect ordering of its cliques.

$$\overset{1}{\bullet} - \overset{2}{\bullet} - \overset{3}{\bullet} - \overset{4}{\bullet}$$

but C_1, C_3, C_2 is not perfect.

We use the notation

$$H_i = \cup_{j=1}^{i} C_j, \quad S_i = C_i \cap (\cup_{j=1}^{i-1} C_j) \quad R_i = C_i \setminus S_i = C_i \setminus H_{i-1}.$$

history separator residual

Here for the perfect order $C_1,\ C_2,\ C_3$,

$$H_1 = C_1,\ H_2 = C_1 \cup C_2,\ S_2 = C_2 \cap C_1 = \{2\},\ H_3 = C_3 \cap (C_1 \cup C_2),\ S_3$$

The Markov property w.r.t. G decomposable

We write $i \sim j$ to indicate i and j are linked.

A distribution is said to be Markov with respect to the graph if

$$X_i \perp X_j | X_{V \setminus \{i,j\}} \text{ whenever } i \not\sim j.$$

For a perfect order of the cliques, $C_1, \ldots, C_k$, we have the following conditional independence relations

$$X_{R_2} \perp X_{C_1} | X_{S_2}, \quad \ldots, \ldots \quad X_{R_k} \perp X_{H_{k-1}} | X_{S_k},$$

General notation, for Σ symmetric

$$\Sigma_{S_i} = \Sigma_{<i>}, \quad \Sigma_{R_i} = \Sigma_{[i]}, \quad \Sigma_{R_i,S_i} = \Sigma_{[i>},$$

$$\Sigma_{R_i \bullet S_i} = \Sigma_{[i]} - \Sigma_{[i>} \Sigma_{<i>}^{-1} \Sigma_{<i]} = \Sigma_{[i]\bullet}$$

The Gaussian model Markov w.r.t. G

Let $C_1, \dots, C_k$ be a perfect order of the cliques.

The normal density factorizes as (see DL93, Th. 2.6)

$$N_X(0, \Sigma) = \frac{\prod_{i=1}^{k} N(0, \Sigma_{C_i})}{\prod_{i=2}^{k} N(0, \Sigma_{S_i})} = \frac{\prod_{C \in \mathcal{C}} N(0, \Sigma_C)}{\prod_{S \in \mathcal{S}} N(0, \Sigma_S)}$$

The parameter is the collection $(\Sigma_C,\ C \in \mathcal{C})$

We have a reduction of the parameter space.

An example $\overset{1}{\bullet} - \overset{2}{\bullet} - \overset{3}{\bullet}$

The conditional covariance between X_3 and X_1 given X_2 is zero, that is

$$\sigma_{31\bullet 2} = \sigma_{31} - \sigma_{32}\sigma_2^{-1}\sigma_{21} = 0$$

and therefore $\hat{\Sigma} = \begin{pmatrix} \sigma_1 & \sigma_{12} & \sigma_{12}\sigma_2^{-1}\sigma_{23} \\ \sigma_{21} & \sigma_2 & \sigma_{23} \\ \sigma_{32}\sigma_2^{-1}\sigma_{21} & \sigma_{32} & \sigma_3 \end{pmatrix}$

The parameter is

$$\Sigma = \begin{pmatrix} \sigma_1 & \sigma_{12} & * \\ \sigma_{21} & \sigma_2 & \sigma_{23} \\ * & \sigma_{32} & \sigma_3 \end{pmatrix} = \left(\Sigma_{(12)},\ \Sigma_{(23)}\right) = \left(\Sigma_{(12)},\ \sigma_{32}\sigma_2^{-1},\ \sigma_{3\cdot 2}\right).$$

while $\Omega = \hat{\Sigma}^{-1} = \begin{pmatrix} \omega_1 & \omega_{12} & 0 \\ \omega_{21} & \omega_2 & \omega_{23} \\ 0 & \omega_{32} & \omega_3 \end{pmatrix}$ because $X_i \perp X_j | X_{V\setminus\{i,j\}} \iff \omega_{ij} = 0$.

The cones P_G and Q_G

- The cone Q_G

Q_G:={incomplete matrices x with missing entries x_{ij}, $(i,j) \notin E$ and such that $x_A > 0$ for $A \subseteq V$ complete}.

- The cone P_G

P_G:={$y \in P^+$ such that $y_{ij} = 0$ whenever $(i,j) \notin E$}.

When G is decomposable, any x in Q_G can be uniquely completed as $\hat{x} \in P^+$ such that for all $(i,j) \in E$

$$x_{ij} = \hat{x}_{ij} \text{ and } \hat{x}^{-1} \in P_G$$

The bijection between P_G and Q_G

Let π denotes the projection of P onto incomplete matrices.

The mapping

$$\varphi \; : \quad y = (\widehat{x})^{-1} \in P_G \mapsto x = \varphi(y) = \pi(y^{-1}) \in Q_G \; ,$$

defines a bijection between P_G and Q_G.

Notation:

$$\text{for } (x, y) \in Q_G \times P_G, \; \text{tr}(xy) = \langle x, y \rangle = \sum_{(i,j) \in E} x_{ij} y_{ij} = \text{tr}(\hat{x} y) \; .$$

Inference for Σ in $\mathcal{N}_G$

$$\mathcal{N}_G = \{N(0,\Sigma) |\, \Omega = \widehat{\Sigma}^{-1} \in P_G\} = \{N(0,\Sigma),\ \ \Sigma \in Q_G\}.$$

- Σ lies in Q_G.
- Ω lies in P_G.
- The Wishart is defined on P^+: not the right cone!.

Question 1: What is the distribution of the MLE of $\Sigma \in Q_G$?

Question 2: Which prior should we put on Σ or what would the induced prior on $\Omega = \hat{\Sigma}^{-1} \in P_G$ be?

The answer is given in DL93!!!

Outline

- The Gaussian model
- The graphical Gaussian model $\mathcal{N}_G$
- The W_{Q_G} and W_{P_G} Wisharts
- The hyper Markov property
- The expected values of W_{Q_G} and IW_{P_G}
- Decision theoretic estimation of Σ
- Another Wishart?

The hyper Wishart for the MLE of Σ

The model $\mathcal{N}_G$ is strong meta Markov:

$$N_X(0,\Sigma) = N_{C_1}(0,\Sigma_{C_1}\ ;\ X_{C_1}) \prod_{i=2}^{k} N_{R_i|S_i}(\Sigma_{R_i,S_i}\Sigma_{S_i}^{-1}x_{S_i},\Sigma_{R_i\bullet S_i}\ ; X_{R_i}|x_{S_i})$$

$\Sigma_{C_1},\ \ (\Sigma_{[i>}\Sigma_{<i>}^{-1},\ \ \Sigma_{[i]\cdot}), i=2,\ldots,k$ are functionally independent.

The marginal model for X_A, for any $A \subseteq V$ complete, is an NEF.

DL (93) show that, then,

the distribution of the MLE of $\Sigma \in Q_G$ is weak hyper Markov.

Therefore......

The hyper Wishart, (cont'd)

The density of the hyper Wishart is therefore

$$W_{Q_G}(p,\sigma;dx) \quad \propto \quad \frac{\prod_{i=1}^{k} w_{c_i}(p,\sigma_{C_i};x_{C_i})}{\prod_{i=2}^{k} w_{s_i}(p,\sigma_{S_i};x_{S_i})} \mathbf{1}_{Q_G}(x)dx$$

with $\quad w_{c_i}(p,\sigma_{C_i};x_{C_i}) = \frac{|x_{C_i}|^{p-\frac{c_i+1}{2}}}{\Gamma_{c_i}(p)|\sigma_{C_i}|^p} e^{-\langle x_{C_i},\sigma_{C_i}^{-1}\rangle}$, that is

$$W_{Q_G}(p,\sigma;dx) \quad \propto \quad \exp -\langle x,\hat{\sigma}^{-1}\rangle \frac{\prod_{i=1}^{k} |x_{C_i}|^{p-\frac{c_i+1}{2}}}{\prod_{2=1}^{k} |x_{S_i}|^{p-\frac{s_i+1}{2}}} \mathbf{1}_{Q_G}(x)dx$$

We note that

1. it is an NEF with only one shape parameter $p = \frac{n}{2}$
2. the expression of $W_{Q_G}(p,\sigma;dx)$ does not depend on the chosen perfect order of the cliques.

The general W_{Q_G} as an NEF on Q_G

We want several shape parameters rather than just p.

The $W_{Q_G}(\alpha, \beta, \sigma)$ family of of distributions (LM 07):

$$W_{Q_G}(\alpha, \beta, \sigma); dx) \quad \propto \quad \frac{\prod_{i=1}^{k} |x_{C_i}|^{\alpha_i - \frac{c_i+1}{2}}}{\prod_{i=2}^{k} |x_{S_i}|^{\beta_i - \frac{s_i+1}{2}}} e^{-\langle x, \hat{\sigma}^{-1} \rangle} \mathbf{1}_{Q_G}(x) dx,$$

is the NEF generated by

$$H_G(\alpha, \beta; x) \mu_G(dy) \quad = \quad \frac{\prod_{i=1}^{k} (\det x_{C_i})^{\alpha_i - \frac{c_i+1}{2}}}{\prod_{i=2}^{k} (\det x_{S_i})^{\beta_i - \frac{s_i+1}{2}}} \, dy$$

The $W_{Q_G}(\frac{n}{2}, \sigma)$, the hyper Wishart, is a special case of the $W_{Q_G}(\alpha, \beta, \theta)$ for $\quad \alpha_i = \frac{n}{2}, \ \ \beta_i = \frac{n}{2}, \ \ \theta = \hat{\sigma}^{-1}.$

The DY conjugate prior distribution for Σ

The hyper inverse Wishart distribution (DL93) on Q_G: a conjugate prior for Σ (in fact induced by the DY prior on Ω).

$$HIW(\frac{\delta+c_i-1}{2}, \frac{\delta+s_i-1}{2}, \theta; dx) \quad \propto \quad \frac{\prod_{i=1}^k |x_{C_i}|^{-\frac{\delta+c_i-1}{2}-\frac{c_i+1}{2}}}{\prod_{i=2}^k |x_{S_i}|^{-\frac{\delta+s_i-1}{2}-\frac{s_i+1}{2}}} e^{-\langle \hat{x}^{-1}, \theta\rangle} \mathbf{1}_{Q_G}(x)dx$$

Same problem: only one shape parameter

The inverse of the hyper inverse Wishart has density

$$\begin{aligned} W_{P_G}(\delta, \theta; dy) \quad &\propto \quad \frac{\prod_{i=1}^k |x_{C_i}(y)|^{-\frac{\delta+c_i-1}{2}+\frac{c_i+1}{2}}}{\prod_{i=2}^k |x_{S_i}(y)|^{-\frac{\delta+s_i-1}{2}+\frac{s_i+1}{2}}} e^{-\langle y, \theta\rangle} \mathbf{1}_{P_G}(y)dy \\ &= \quad |y|^{\frac{\delta-2}{2}} e^{-\langle y, \theta\rangle} \mathbf{1}_{P_G}(y)dy. \end{aligned}$$

The $W_{P_G}(\delta, \theta; dy)$ is a .....natural exponential family.

The general W_{P_G} as a NEF on P_G

The $W_{P_G}(\alpha, \beta, D)$ family of of distributions (LM 07)

$$W_{P_G}(\alpha, \beta, \theta); dy) \quad \propto \quad \frac{\prod_{i=1}^{k} |x_{C_i}(y)|^{\alpha_i + \frac{c_i+1}{2}}}{\prod_{i=2}^{k} |x_{S_i}(y)|^{\beta_i + \frac{s_i+1}{2}}} e^{-\langle y, \theta \rangle} \mathbf{1}_{P_G}(y) dy,$$

is the NEF generated by

$$H_G(\alpha, \beta; x(y))\nu_G(dy) \quad = \quad \frac{\prod_{i=1}^{k} (\det x_{C_i}(y))^{\alpha_i + \frac{c_i+1}{2}}}{\prod_{i=2}^{k} (\det x_{S_i}(y))^{\beta_i + \frac{s_i+1}{2}}} \; dy$$

The $W_{P_G}(\delta, \theta; dy)$, inverse of the HIW, is a special case of the $W_{P_G}(\alpha, \beta, \theta)$ for $\alpha_i = -\frac{\delta + c_i - 1}{2}, \quad \beta_i = -\frac{\delta + s_i - 1}{2}.$

The $IW_{P_G}(\alpha, \beta, \theta)$ as a prior

The inverse W_{P_G} is the $IW_{P_G}(\alpha, \beta, \theta)$, defined on Q_G

$$IW_{P_G}((\alpha, \beta), \theta; d\Sigma)$$

$$= \frac{1}{\Gamma_{II}(\alpha, \beta)} \frac{|\theta_{C_i}|^{\alpha_i}}{|\theta_{S_i}|^{\beta_i}} \times \frac{\prod_{i=1}^{k} |\Sigma_{C_i}|^{\frac{\alpha_i}{2} - \frac{c_i+1}{2}}}{\prod_{i=2}^{k} |\Sigma_{S_i}|^{\frac{\beta_i}{2} - \frac{s_i+1}{2}}} e^{-\langle \hat{\Sigma}^{-1}, \theta \rangle} \mathbf{1}_{Q_G}(\Sigma) d\Sigma$$

Clearly if $Z_i \sim N(0, \Sigma) \in \mathcal{N}_G$ and we write $U = \sum_{i=1}^{n} Z_i Z_i^t$

$$\prod_{i=1}^{n} N(0, \Sigma; dZ_i) IW_{P_G}((\alpha, \beta), \theta; d\Sigma)$$

$$\propto \frac{\prod_{i=1}^{k} |\Sigma_{C_i}|^{\frac{\alpha_i - n}{2} - \frac{c_i+1}{2}}}{\prod_{i=2}^{k} |\Sigma_{S_i}|^{\frac{\beta_i - n}{2} - \frac{s_i+1}{2}}} e^{-\langle \hat{\Sigma}^{-1}, \theta + U \rangle} \mathbf{1}_{Q_G}(d\Sigma) \prod_{i=1}^{k} dZ_i$$

The IW_{P_G} is a conjugate prior for $\Sigma \in Q_G$.

The parameter sets for the W_{Q_G} and W_{P_G}

For $\sigma \in Q_G$ and $\theta \in Q_G$, let

$$\mathcal{A} = \left\{ (\alpha, \beta) \ : \ \int_{Q_G} e^{-\langle x, \hat{\sigma}^{-1} \rangle} H_G(\alpha, \beta; x) \mu_G(dx) = \Gamma_I(\alpha, \beta) H_G(\alpha, \beta; \sigma) \right\}$$

and

$$\mathcal{B} = \left\{ (\alpha, \beta) \ : \ \int_{P_G} e^{-\langle y, \theta \rangle} H_G(\alpha, \beta; \varphi(y) \nu_G(dy) = \Gamma_{II}(\alpha, \beta) H_G(\alpha, \beta; \theta) \right\}$$

where $\Gamma_I(\alpha, \beta)$ and $\Gamma_{II}(\alpha, \beta)$ are functions of (α, β) only,

We define

- the $W_{Q_G}(\alpha, \beta, \sigma; dx)$ for $(\alpha, \beta) \in \mathcal{A}, \ \sigma \in Q_G$
- the $W_{P_G}(\alpha, \beta, \theta; dy)$ for $(\alpha, \beta) \in \mathcal{B}, \ \theta \in Q_G$.

Outline

- The Gaussian model
- The graphical Gaussian model $\mathcal{N}_G$
- The W_{Q_G} and W_{P_G} Wisharts
- The hyper Markov property
- The expected values of W_{Q_G} and IW_{P_G}
- Decision theoretic estimation of Σ
- Another Wishart?

The hyper Markov property

Recall that

- the IW_{P_G} is a conjugate prior for $\Sigma \in Q_G$ in $\mathcal{N}_G$
- the HIW is the DY conjugate prior for $\Sigma \in Q_G$ in $\mathcal{N}_G$
- $$N_X(0,\Sigma) \;=\; N_{X_{C_1}}(0,\Sigma_{C_1})\prod_{i=2}^{k} N_{X_{R_i}|X_{S_i}}(\Sigma_{R_i,S_i}\Sigma_{S_i}^{-1}x_{S_i},\Sigma_{R_i\bullet S_i})$$
- the HIW is strong hyper Markov, i.e., under the HIW

$$\Sigma_{C_1},\ \perp \left(\Sigma_{[i>}\Sigma_{<i>}^{-1},\Sigma_{[i]\cdot}\right),\, i=2,\ldots,k$$

To show the hyper Markov property, we change variables

$$\Sigma \in Q_G \mapsto \left(\Sigma_{C_1},\ \left(\Sigma_{[i>}\Sigma_{<i>}^{-1},\Sigma_{[i]\cdot}\right),\, i=2,\ldots,k\right)$$

The normalizing constant for the W_{P_G}

We work in Q_G.

$$\int_{Q_G} e^{-\langle \theta, \hat{x}^{-1}\rangle} \frac{\prod_{j=1}^k |x_{C_j}|^{\alpha_j - \frac{c_j+1}{2}}}{\prod_{S\in\mathcal{S}} |x_S|^{\nu(S)(\beta(S)-\frac{|S|+1}{2})}} dx$$

$$= \int |x_{C_1}|^{\alpha_1 - \frac{c_1+1}{2}} e^{-\langle x_{C_1}^{-1}, \theta_{C_1}\rangle} \prod_{j=2}^k |x_{[j].}|^{\alpha_j - \frac{c_j+1}{2}} e^{-\langle x_{[j].}^{-1}, \theta_{[j].}\rangle}$$

$$\prod_{j=2}^k e^{-\langle (x_{[j>} x_{<j>}^{-1} - \theta_{[j>}\theta_{<j>}^{-1}), x_{[j].}^{-1}(x_{[j>} x_{<j>}^{-1} - \theta_{[j>}\theta_{<j>}^{-1})\theta_{<j>}\rangle}$$

$$\prod_{S\in\mathcal{S}} |x_S|^{\sum_{i\in J(P,S)}(\alpha_i - \frac{c_i+1}{2}) - \nu(S)(\beta(S)-\frac{|S|+1}{2})}$$

$$\prod_{S\in\mathcal{S}} |x_S|^{\sum_{i\in J(P,S)} c_i - \nu(S)|S|} dx_{C_1} \prod_{j=2}^k d(x_{[j>} x_{<j>}^{-1}) dx_{[j].}.$$

The strong Hyper Markov property

- For the HIW, i.e., when $\alpha_i = \frac{\delta + c_i - 1}{2}, \beta_i = \frac{\delta + s_i - 1}{2}$

The terms in red disappear!

And we obtain the normalizing constant *and* the strong hyper Markov property,

- For the general W_{P_G}, the terms do not disappear unless we choose (α, β) carefully.

This choice *depends upon the chosen perfect order of the cliques,* P:
For $(\alpha, \beta) \in B_P$, we obtain

1. the normalizing constant $H_G(\alpha, \beta; \theta)\Gamma_{II}(\alpha, \beta)$
2. the strong directed hyper Markov property.

The set B_P

For a given perfect order P of the cliques, B_P is the set of (α, β) such that

1. $\sum_{j \in J(P,S)} (\alpha_j + \frac{1}{2}(c_j - s_j)) - \nu(S)\beta(S) = 0$ for all S different from S_2;
2. $-\alpha_q - \frac{1}{2}(c_q - s_q - 1) > 0$ for all $q = 2, \ldots, k$ and $-\alpha_1 - \frac{1}{2}(c_1 - s_2 - 1) > 0$
3. $-\alpha_1 - \frac{1}{2}(c_1 - s_2 + 1) - \gamma_2 > \frac{s_2 - 1}{2}$ where $\gamma_2 = \sum_{j \in J(P,S_2)} \left(\alpha_j - \beta_2 + \frac{c_j - s_2}{2}\right)$.

a set of linear constraints that reduces the number of free parameters to $k+1$: $\beta_2,\ \alpha_i,\ i = 1, \ldots, k$

$$\mathcal{B} \supseteq \cup_P B_P$$

The strong directed hyper Markov property

If $\Omega \sim W_{P_G}(\alpha, \beta, \theta)$, i.e. $\Sigma = \varphi(Y) \sim IW_{P_G}(\alpha, \beta, \theta)$ with $(\alpha, \beta) \in B_P$ and $\theta \in Q_G$, then

$$\begin{aligned}
\Sigma_{[12>}|\Sigma_{[1]\cdot} &\sim N_{(c_1-s_2)\times s_2}(\theta_{[12>}, 2\,\theta^{-1}_{<2>} \otimes \Sigma_{[1]\cdot}) \\
\Sigma_{<2>} &\sim iw_{s_2}(-(\alpha_1 + \frac{c_1 - s_2}{2} + \gamma_2), \theta_{<2>}) \\
\Sigma_{[i]\cdot} &\sim iw_{c_i - s_i}(-\alpha_i,\ \theta_{[i]\cdot}),\quad i = 1, \dots, k \\
\Sigma_{[j>}\Sigma^{-1}_{<j>}|\Sigma_{[j]\cdot} &\sim N_{(c_j-s_j)\times s_j}(\theta_{[j>}\theta^{-1}_{<j>}, 2\,\theta^{-1}_{<j>} \otimes x_{[j]\cdot}),\quad j = 2, \dots,
\end{aligned}$$

and

$$\{(\Sigma_{[12>}, \Sigma_{[1]\cdot}),\ \Sigma_{<2>}, (\Sigma_{[j>}\Sigma^{-1}_{<j>}, \Sigma_{[j]\cdot}), j = 2, \dots, k\} \qquad (3)$$

are mutually independent.

The IW_{P_G} is strong directed hyper Markov.

An improper member of the IW_{P_G} family

Let $\sigma = 2\Sigma$ and let ϕ be its Choleski parametrization:

$$\phi = (\sigma_{[1]\cdot}^{-1}, \sigma_{[12>}, \sigma_{<2>}, \sigma_{[j]\cdot}^{-1}, \sigma_{[j>}\sigma_{<j>}^{-1}, j = 2, \ldots, k)$$

Since the the hyper Wishart is an NEF, we can use the method of Data and Ghosh (1995) to obtain the reference prior for the parameter $\sigma \in Q_G$ as

$$\pi^{\sigma}(\sigma) \quad \propto \quad \frac{|\sigma_{C_1}|^{-\frac{c_1+1}{2}} \prod_{j=2}^{k} |\sigma_{C_j}|^{-\frac{c_j+1}{2}}}{|\sigma_{S_2}|^{\frac{c_1+c_2}{2}-s_2-\frac{s_2+1}{2}} \prod_{j=3}^{k} |\sigma_{S_j}|^{\frac{c_j-s_j}{2}-\frac{s_j+1}{2}}}$$

It is an improper $IW_{P_G}(\alpha, \beta, 0; x)$ distribution with

$$\alpha_j = 0,\ j = 1, \ldots, k,\quad \beta_2 = \frac{c_1 + c_2}{2} - s_2, \beta_j = \frac{c_j - s_j}{2},\ j = 3, \ldots,$$

No need to choose hyperparameters here

The W_{Q_G} is weak hyper Markov

If $X \sim W_{Q_G}(\alpha, \beta, \sigma)$ with $(\alpha, \beta) \in A_P$ and $\sigma \in Q_G$,
then

$$
\begin{aligned}
x_{[1]\cdot} &\sim w_{c_1 - s_2}(\alpha_1 - \frac{s_2}{2}, \sigma_{[1]\cdot}) \\
x_{[12>} | x_{<2>} &\sim N_{(c_1 - s_2) \times s_2}(\sigma_{[12>}, 2\, x_{<2>}^{-1} \otimes \sigma_{[1]\cdot}) \\
x_{<2>} &\sim w_{s_2}(\alpha_1 + \delta_2, \sigma_{<2>}) \\
x_{[j>} x_{<j>}^{-1} | x_{<j>} &\sim N_{(c_j - s_j) \times s_j}(\sigma_{[j>} \sigma_{<j>}^{-1}, 2\, x_{<j>}^{-1} \otimes \sigma_{[j]\cdot}) \\
x_{[j]\cdot} &\sim w_{c_j - s_j}(\alpha_j - \frac{s_j}{2}, \sigma_{[j]\cdot}), j = 2, \ldots, k
\end{aligned}
$$

The distribution of $x_{[j>} x_{<j>}^{-1}$ depends upon $x_{<j>}$.

Outline

- The Gaussian model
- The graphical Gaussian model $\mathcal{N}_G$
- The W_{Q_G} and W_{P_G} Wisharts
- The hyper Markov property
- The expected values of W_{Q_G} and IW_{P_G}
- Decision theoretic estimation of Σ
- Another Wishart?

Expected values

For the estimation of the covariance Σ,

using the IW_{P_G} as a prior, we will need

$$E(W_{P_G})^{-1} \quad \text{and} \quad E(IW_{P_G}).$$

These are explicit expressions.

No need for MCMC computations.

The inverse of $E(W_{P_G})$, i.e. $E(\Omega|S)^{-1}$

$$
\begin{aligned}
&[E(\Omega|S)] \\
&= \left[E(W_{P_G}(\alpha_j - \frac{n}{2}, \beta_j - \frac{n}{2}, \theta + \kappa(nS))\right] \\
&= -\frac{1}{2}\left[\sum_{j=1}^{k}(\alpha_j - \frac{n}{2})(\theta + \kappa(nS))_{C_j}^{-1})^0 - \sum_{j=2}^{k}(\beta_j - \frac{n}{2})(\theta + \kappa(nS))_{S_j}^{-1})^0\right]
\end{aligned}
$$

explicit analytic expression: no recourse to MCMC

The $E(IW_{P_G})$ computed layer by layer

Let $\xi = \theta + nS$

$$E(X_{<2>}) = \frac{\xi_{<2>}}{-(\alpha_1 + \frac{c_1 - s_2}{2} + \gamma_2) - \frac{s_2+1}{2}} = \frac{\xi_{<2>}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

explicit analytic expression: computed sequentially

The $E(IW_{P_G})$ computed layer by layer

Let $\xi = \theta + nS$

$$E(X_{<2>}) = \frac{\xi_{<2>}}{-(\alpha_1 + \frac{c_1 - s_2}{2} + \gamma_2) - \frac{s_2+1}{2}} = \frac{\xi_{<2>}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

$$E(X_{C_1 \setminus S_2, S_2}) = \frac{\xi_{C_1 \setminus S_2, S_2}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

explicit analytic expression: computed sequentially

The $E(IW_{P_G})$ computed layer by layer

Let $\xi = \theta + nS$

$$E(X_{<2>}) = \frac{\xi_{<2>}}{-(\alpha_1 + \frac{c_1 - s_2}{2} + \gamma_2) - \frac{s_2+1}{2}} = \frac{\xi_{<2>}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

$$E(X_{C_1 \setminus S_2, S_2}) = \frac{\xi_{C_1 \setminus S_2, S_2}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

$$E(X_{C_1 \setminus S_2}) = \frac{\xi_{[1]\cdot}}{-(\alpha_1 + \frac{c_1 - s_2 + 1}{2})} \left(1 - \frac{s_2}{2(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}\right) + \frac{\xi_{C_1 \setminus S_2, S_2} \xi_{<2>}^{-1} \xi_{S_2, C_1 \setminus S_2}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

explicit analytic expression: computed sequentially

The $E(IW_{P_G})$ computed layer by layer

Let $\xi = \theta + nS$

$$E(X_{<2>}) = \frac{\xi_{<2>}}{-(\alpha_1 + \frac{c_1 - s_2}{2} + \gamma_2) - \frac{s_2+1}{2}} = \frac{\xi_{<2>}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

$$E(X_{C_1 \setminus S_2, S_2}) = \frac{\xi_{C_1 \setminus S_2, S_2}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

$$E(X_{C_1 \setminus S_2}) = \frac{\xi_{[1]\cdot}}{-(\alpha_1 + \frac{c_1 - s_2 + 1}{2})}\left(1 - \frac{s_2}{2(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}\right) + \frac{\xi_{C_1 \setminus S_2, S_2}\xi^{-1}_{<2>}\xi_{S_2, C_1 \setminus S_2}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

$$E(X_{R_j, S_j}) = \xi_{[j>}\xi^{-1}_{<j>}E(X_{<j>}), \qquad j = 2, \ldots, k$$

explicit analytic expression: computed sequentially

The $E(IW_{P_G})$ computed layer by layer

Let $\xi = \theta + nS$

$$E(X_{<2>}) = \frac{\xi_{<2>}}{-(\alpha_1 + \frac{c_1 - s_2}{2} + \gamma_2) - \frac{s_2+1}{2}} = \frac{\xi_{<2>}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

$$E(X_{C_1 \setminus S_2, S_2}) = \frac{\xi_{C_1 \setminus S_2, S_2}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

$$E(X_{C_1 \setminus S_2}) = \frac{\xi_{[1]\cdot}}{-(\alpha_1 + \frac{c_1 - s_2 + 1}{2})} \left(1 - \frac{s_2}{2(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}\right) + \frac{\xi_{C_1 \setminus S_2, S_2} \xi^{-1}_{<2>} \xi_{S_2, C_1 \setminus S_2}}{-(\alpha_1 + \frac{c_1+1}{2} + \gamma_2)}$$

$$E(X_{R_j, S_j}) = \xi_{[j>} \xi^{-1}_{<j>} E(X_{<j>}), \qquad j = 2, \ldots, k$$

$$E(X_{R_j}) = \frac{\xi_{[j]\cdot}}{-(\alpha_j + \frac{c_j - s_j + 1}{2})} \left(1 + \frac{1}{2} \text{tr}(\xi^{-1}_{<j>} E(X_{<j>}))\right) + \xi_{[j>} \xi^{-1}_{<j>} E(X_{<j>}) \xi^{-1}_{<j>} \xi_{<j]}$$

explicit analytic expression: computed sequentially

Outline

- The Gaussian model
- The graphical Gaussian model $\mathcal{N}_G$
- The W_{Q_G} and W_{P_G} Wisharts
- The hyper Markov property
- The expected values of W_{Q_G} and IW_{P_G}
- Decision theoretic estimation of Σ
- Another Wishart?

The loss functions

Loss functions for Σ and Ω

$$L_1(\tilde{\Sigma}) = \text{tr}(\tilde{\Sigma}\widehat{\Sigma}^{-1}) - \log(|\tilde{\Sigma}\widehat{\Sigma}^{-1}|) - r \qquad L_2(\tilde{\Sigma}) = \text{tr}(\tilde{\Sigma} - \Sigma)^2$$
$$L_1(\tilde{\Omega}) = \text{tr}(\tilde{\Omega}\Omega^{-1}) - \log(|\tilde{\Omega}\Omega^{-1}|) - r \qquad L_2(\tilde{\Omega}) = \text{tr}(\tilde{\Omega} - \Omega)^2$$

Can we use them as is? It is important to note that

$$L_1(\tilde{\Sigma}) = \sum_{C\in\mathcal{C}} \text{tr}\tilde{\Sigma}_C\Sigma_C^{-1} - \sum_{S\in\mathcal{S}} \text{tr}\tilde{\Sigma}_S\Sigma_S^{-1} - \log \frac{\prod_{C\in\mathcal{C}} |\tilde{\Sigma}_C| \prod_{S\in\mathcal{S}} |\Sigma_S|}{\prod_{C\in\mathcal{C}} |\Sigma_C| \prod_{S\in\mathcal{C}} |\tilde{\Sigma}_S|}$$
$$L_2(\tilde{\Sigma}) = \Sigma_{(i,j)\in E}(\tilde{\Sigma}_{ij} - \Sigma_{ij})^2$$

Similarly $L_1(\tilde{\Omega}), L_2(\tilde{\Omega})$ only use the non zero entries of Ω

Our estimators

The Bayes estimators under L_1, L_2 and the IW_{P_G} on $\Sigma \in Q_G$ (equivalently the W_{P_G} on $\Omega \in P_G$)

Parameter of interest	L1	L2
Σ	$\tilde{\Sigma}_{L_1} = [E(\Omega\|S)]^{-1}$	$\tilde{\Sigma}_{L_2} = E(\Sigma\|S)$
Ω	$\tilde{\Omega}_{L_1} = [E(\Sigma\|S)]^{-1}$	$\tilde{\Omega}_{L_2} = E(\Omega\|S)$

The risk functions

Duality: The relationship between our Bayes estimators is as follows

$$\tilde{\Sigma}_{L_1} = \pi\left([\tilde{\Omega}_{L_2}]^{-1}\right)$$
$$\tilde{\Sigma}_{L_2} = \pi\left([\tilde{\Omega}_{L_1}]^{-1}\right)$$

The risk functions We assess the quality of our estimators using risk comparison for:

$$R_{L_i}(\tilde{\Sigma}_{L_i}) = E[L_i(\tilde{\Sigma}_{L_i}, \Sigma)], \quad i = 1, 2$$
$$R_{L_i}(\tilde{\Omega}_{L_i}) = E[(L_i(\tilde{\Omega}_{L_i}, \Omega)], \quad i = 1, 2$$

The prior, loss functions and estimators

We will use three different priors for Σ

- The $IW_{P_G}(\alpha, \beta, \theta)$ with $k+1$ free shape parameters
- The $HIW(\delta, \theta)$ prior with 1 free shape parameter: a special case of the IW_{P_G}
- The reference prior: an objective prior

and two loss functions L_1 and L_2, and four estimators

$$E(\Omega|S) \text{ and its inverse } [E(\Omega|S)]^{-1}$$

$$E(\Sigma|S) \text{ and its inverse } [E(\Sigma|S)]^{-1}$$

and the MLE and the $MLEg$, that is the mle under the graphical model.

So, we have a total of eight estimators that we are going to study and compare.

Our approach: Bayesian graphical models

Bayesian graphical models combines the two approaches

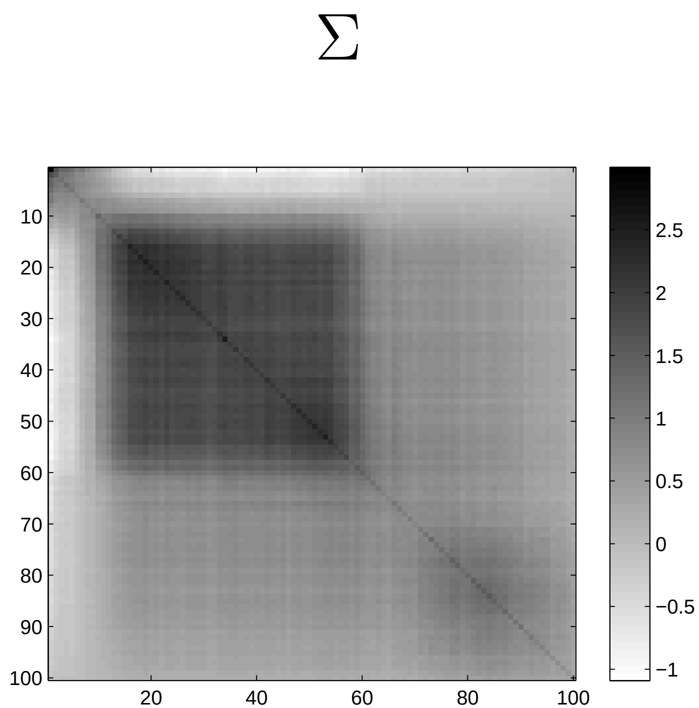
- The graphical model is used as a tool for regularization
- The prior give us flexibility in the estimator

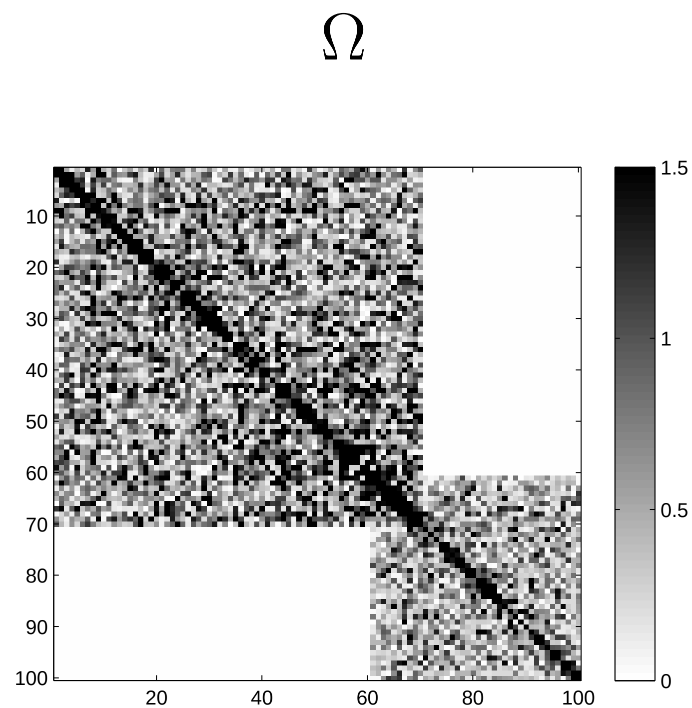
Traditional choice of priors

- The conjugate prior which is the Wishart $W_r(\delta, \theta)$ with one shape parameter δ and the scale parameter θ
- Various priors which give more flexibility for the parameters, (inverse gammas on the diagonal and independent normals on the triangular elements of the Choleski) but then you

lose conjugacy- - - problematic for computations in high-dimensional problems.

"Two Cliques" study





Simple Example with 2 Cliques

$p = 100$, and $n = 75, 100, 500, 1000$, C1=70 and C2=40

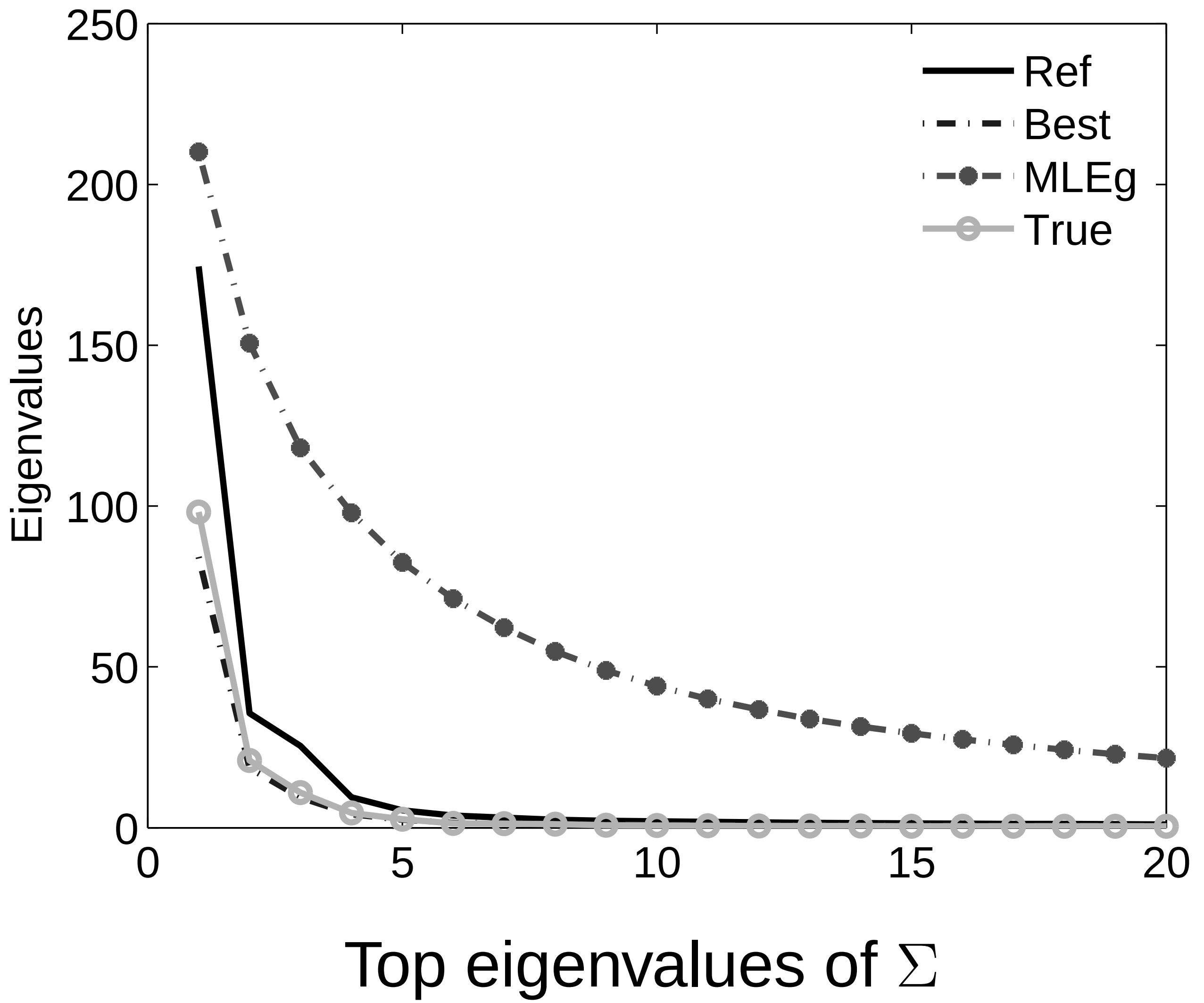
Scale Hyperparameter : $\theta = I$ or θ s.t. prior expected value of $\Sigma \in Q_G$ is I

Goal: Explore the flexibility of the IWpg

"Two Cliques" study - Risk comparis

	$n = 75$		$n = 100$		$n = 500$		$n = 10$	
	$R_1(\Omega)$	$R_1(\Sigma)$	$R_1(\Omega)$	$R_1(\Sigma)$	$R_1(\Omega)$	$R_1(\Sigma)$	$R_1(\Omega)$	
Reference	212.7	66.61	60.71	40.93	7.02	6.66	3.33	
$HIW(3, I)$	98.76	59.28	80.72	43.41	7.76	7.18	3.54	
$IW_{P_G}(1/2c_i, D)$	29.99	25.18	24.53	24.49	6.37	6.21	3.27	
$IW_{P_G}(1/2c_i, I)$	207.4	67.88	116.7	49.78	8.69	7.80	3.76	
$IW_{P_G}(1/4c_i, D)$	red22.18	red 17.96	red18.57	red15.87	red5.67	red5.43	red3.03	r
$IW_{P_G}(1/4c_i, I)$	165.5	63.10	96.14	46.20	8.14	7.43	3.67	
$IW_{P_G}(1/10c_i, D)$	35.71	31.99	31.59	27.02	6.77	6.41	3.32	
$IW_{P_G}(1/10c_i, I)$	141.7	60.23	89.67	45.03	7.98	7.32	3.59	
MLEg	813.9	70.72	154.6	43.51	8.13	6.79	3.62	
MLE	–	–	7.3×10^8	102.5	14.45	10.85	6.00	
Risk Reduc. vs MLEg	red97%	red 75%	red88%	red64%	red30%	red20%	red16%	r

"Two Cliques" study - Scree Plots



Call-center data: Fitting a Graphical model

Dataset also analyzed by Huang et al. (2006) and Bickel and Levina(2006)

Records from a call-center of a major financial institution

Phone call from 7:00am till midnight during 2002 (on week days only)

Recording period of 10-minute intervals during 17 hours

Number of calls in each period N_{ij} , $i = 1, 2, ..239$ and $j = 1, 2, .., 102$ was recorded.

Standard transformation $x_{ij} = \left(N_{ij} + \frac{1}{4}\right)^{\frac{1}{2}}$ applied to make data closer to Normal

First 205 data points as training data and remaining 34 as test data

Call-center data: Fitting a Graphical model

Aim: Choose the “best" graphical model for the data, among models with banded inverse covariance matrices.

Criterion 1: K-fold cross-validation error ($K = 10$).
We predict the second half of day given first half for the test data set after training our estimators on the training data set

$$x = \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{pmatrix}$$

The best linear predictor for $x_i^{(2)}$ from $x_i^{(1)}$ is

$$x_i^{(2)} = \mu_2 + \Sigma_{21}\Sigma_1^{-1}(x_i^{(1)} - \mu_1)$$

Call-center data: Fitting a Graphical model..

Criteria 2: Bayesian model selection

Maximizing posterior probabilities for the model: we choose G_k with maximum posterior probability i.e.

$$P(G_k|y) \propto p(y|G_k)p(G_k)$$

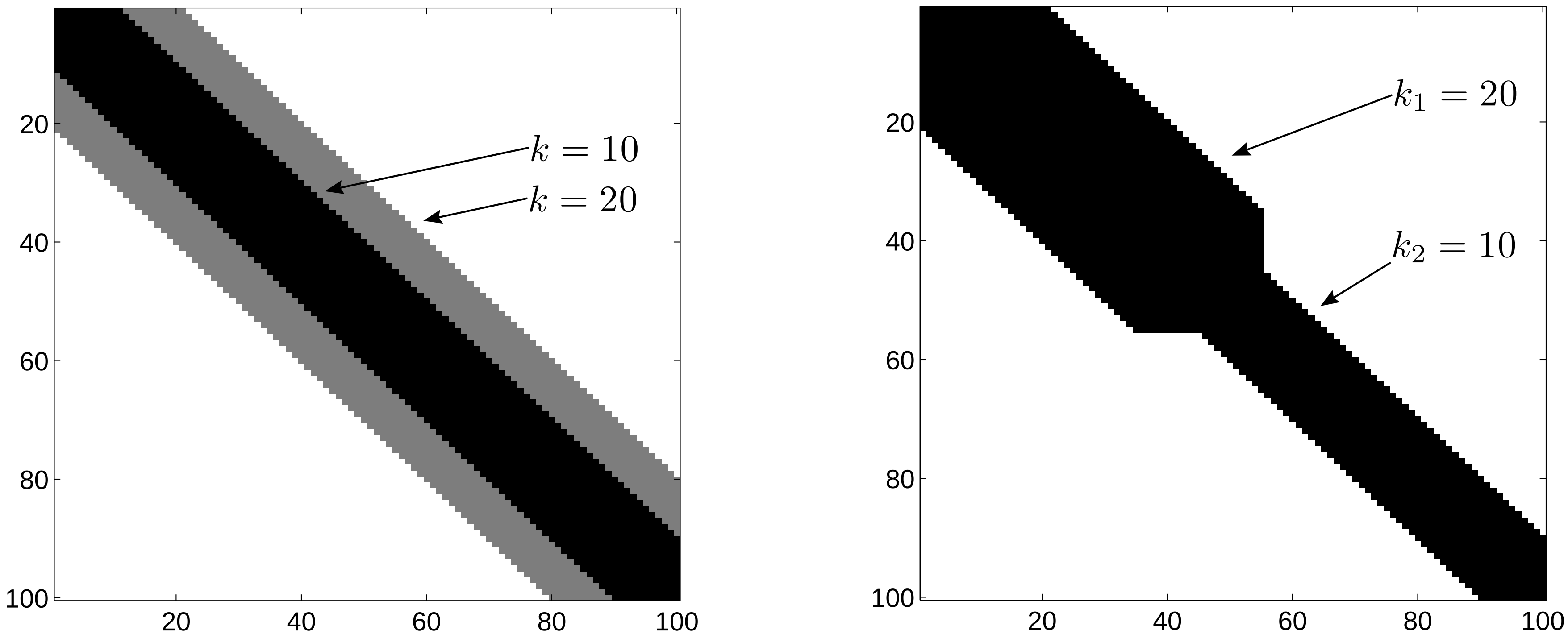
where

$$P(y|G_k) = \int N(y|\Sigma_k) IW_{P_G}(\alpha, \beta, \theta;\ \Sigma_k)\ d\Sigma_k$$

...with some abuse of notation.

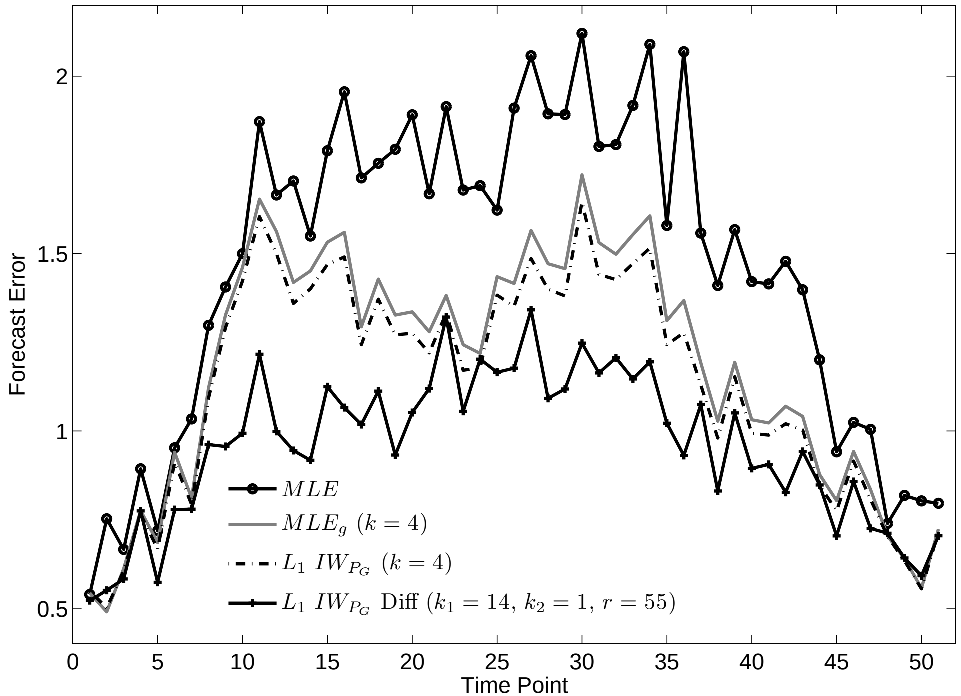
In fact $P(y, G_k)$ is equal to the ratio of normalizing constants for the prior and posterior distributions and these are known explicitly for the IW_{P_G}.

Differential banding



Differential banding illustration

Prediction error



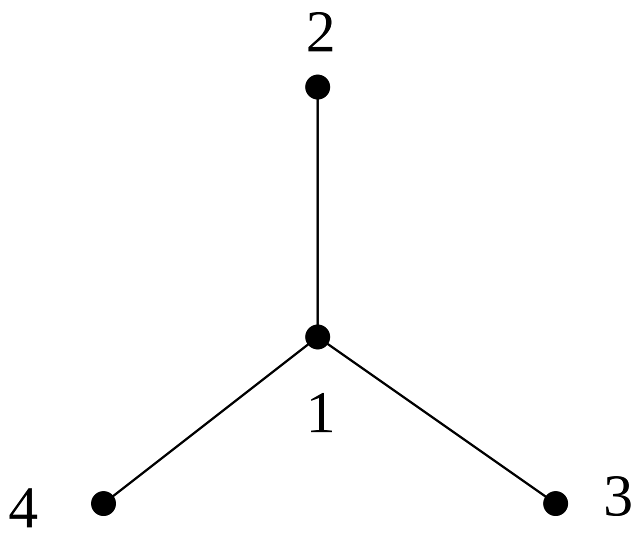
Forecast error for selected banded and “differentially banded” models

Outline

- The Gaussian model
- The graphical Gaussian model $\mathcal{N}_G$
- The W_{Q_G} and W_{P_G} Wisharts
- The hyper Markov property
- The expected values of W_{Q_G} and IW_{P_G}
- Decision theoretic estimation of Σ
- Another Wishart?

The MLE for missing data

Question: Is the W_{Q_G} the distribution of the MLE of Σ?



The notation

$$\mathbf{Z}^t = (Z_1^t, Z_2^t, Z_3^t, Z_4^t)$$
$$\mathbf{Z_i}^t = (Z_1^t, Z_i^t),\ i = 2, 3, 4$$

Z_i is a $p_i \times 1$ vector

The data

$$\mathbf{Z}_1, \dots, \mathbf{Z}_n, \qquad (Z_1)_1, \dots, (Z_1)_{n_1}, \qquad \mathbf{Z_i}_1, \dots, \mathbf{Z_i}_{n_i},\ i = 2, 3, 4$$

$$V_0 = \sum_{j=1}^{n} \mathbf{Z}_j \mathbf{Z}_j^t, \quad V_1 = \sum_{j=1}^{n_1} (Z_1)_j (Z_1)_j^t, \quad V_i = \sum_{j=1}^{n_i} \mathbf{Z_i}_j \mathbf{Z_i}_j^t,\ i = 2, 3, 4$$

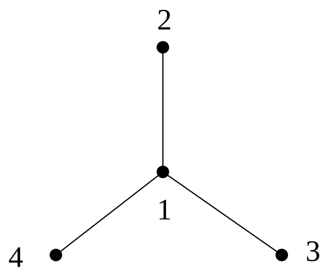
The MLE for missing data: notation

$$v_0 = (v_{0lm},\ l,m = 1,\dots k),\quad v_i = \begin{pmatrix} v_{i11} & v_{i1i} \\ v_{ii1} & v_{iii} \end{pmatrix}, i = 1,2,3,4$$

$$v_{(0i)} = \begin{pmatrix} v_{011} & v_{01i} \\ v_{0i1} & v_{0ii} \end{pmatrix},\ w_i = \begin{pmatrix} v_{011} + v_{i11} & v_{01i} + v_{i1i} \\ v_{0i1} + v_{ii1} & v_{0ii} + v_{iii} \end{pmatrix}, i = 2,3,4$$

$$\begin{aligned} \rho_{i1} &= (v_{0i1} + v_{i21})(v_{011} + v_{i11})^{-1} = \rho_{1i}^t \\ w_{i\cdot 1} &= v_{0ii} + v_{i22} - \rho_{i1}(v_{011} + v_{i11})\rho_{1i} \\ s_1 &= v_{011} + v_1 + \sum_{i=2}^{k} v_{i11}\ . \end{aligned}$$

The MLE: Sun and Sun 07



Based on the incomplete data as given above, the maximum likelihood estimate $\hat{\Sigma}$ of $\Sigma \in Q_G$ is given by the elements of its Choleski decomposition as follows

$$
\begin{aligned}
\hat{\Sigma}_{11} &= \frac{s_1}{m_1} \\
\hat{\Sigma}_{i1}\hat{\Sigma}_{11}^{-1} &= \widehat{\Sigma_{i1}\Sigma_{11}^{-1}} = \rho_{i1},\ i = 2, \ldots, k \\
\widehat{\Sigma_{ii\cdot 1}} &= \frac{w_{i\cdot 1}}{m_i},\ i = 2, \ldots, k
\end{aligned}
$$

We want the joint distribution of $(s_1, \rho_i, w_{i\cdot 1}, i = 2, 3, 4)$

The W_{Q_G} for our graph

$$W^*_{Q_G}(\alpha,\beta,\sigma;\ ds_1,\ d\rho_{i1},w_{i\cdot 1},\ \ i=,2,3,4) \tag{4}$$

$$\propto \quad |s_1|^{\frac{n+n_1+\sum_{i=2}^k n_i}{2}-\frac{p_1+1}{2}} \exp -\frac{1}{2}\langle s_1,\Sigma_{11}^{-1}\rangle$$

$$\times \prod_{i=2}^{4} |s_1|^{\frac{p_i}{2}} \exp -\frac{1}{2}\langle(\rho_{i1}-\Sigma_{i1}\Sigma_{11}^{-1}),\Sigma_{i\cdot 1}^{-1}(\rho_{i1}-\Sigma_{i1}\Sigma_{11}^{-1})^t s_1\rangle$$

$$\times \prod_{i=2}^{k} |w_{i\cdot 1}|^{\frac{n+n_i-p_i}{2}-\frac{p_i+1}{2}} \exp -\frac{1}{2}\langle w_{i\cdot 1},\Sigma_{i\cdot 1}^{-1}\rangle$$

The MLE:the ingredients

$$w_0 = \pi(v_0) = \begin{pmatrix} v_{011} & v_{012} & v_{013} & v_{014} \\ v_{021} & v_{022} & * & * \\ v_{031} & * & v_{033} & * \\ v_{041} & * & * & v_{044} \end{pmatrix} \sim W_{Q_G}(\frac{n}{2}, \frac{n}{2}, \hat{\Sigma}^{-1})$$

$$v_1 \sim W_{p_1}(\frac{n_1}{2}, \Sigma_{11}), \quad w_i \sim W_{p_1+p_i}(\frac{n+n_i}{2}, \Sigma_{(1i)}), \; i = 2, 3, 4$$

Recall $v_0, v_1, v_i, i = 2, 3, 4$ are independent

BUT

the w_i's are NOT independent. They have v_{011} in common.

A few Jacobians later

$$
\begin{aligned}
& f(s_1, \rho_i, w_{i\cdot 1}, i = 2, \ldots, k) \\
\propto \quad & |s_1|^{\frac{n+n_1+\sum_{i=2}^{k}(n_i+p_i)}{2} - \frac{p_1+1}{2}} \exp -\frac{1}{2}\langle s_1, \Sigma_{11}^{-1}\rangle \\
& \times \int_{\mathcal{D}} |l_0|^{\frac{n}{2} - \frac{p_1+1}{2}} |I_{p_1} - l_0 - \sum_{i=2}^{k} l_i|^{\frac{n_1}{2} - \frac{p_1+1}{2}} \prod_{i=2}^{k} |l_i|^{\frac{n_i}{2} - \frac{p_1+1}{2}} \\
& \times \prod_{i=2}^{k} |l_0 + l_i|^{\frac{p_i}{2}} \exp -\frac{1}{2}(\rho_{i1} - \xi_{i1}), \xi_{i\cdot 1}^{-1}(\rho_{i1} - \xi_{i1})^t \sigma(l_0 + l_i)\sigma^t \\
& \times dl_0 \prod_{i=2}^{k} dl_i \times \prod_{i=2}^{k} |w_{i\cdot 1}|^{\frac{n+n_i-p_i}{2} - \frac{p_i+1}{2}} \exp -\frac{1}{2}\langle w_{i\cdot 1}, \Sigma_{i\cdot 1}^{-1}\rangle
\end{aligned}
$$

where $s_1 = \sigma\sigma^t$, σ is a lower triangular matrix.

Another Wishart?

Nearly the W_{Q_G} but not the W_{Q_G}!!