Bayesian inference for covariance matrix

Ignacio Alvarez

IESTA, Universidad de la Repblica, Uruguay

ISEC 2018

Introduction

Covariance matrix estimation

- Multivariate normal sampling models
- random-intercept, random-slope models

$$
\begin{aligned}
y_{ij} &= \beta_{0j} + \beta_{1j}x_{ij} + \beta_{2j}z_{ij} + \varepsilon_{ij} \\
\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \end{pmatrix} &\sim N\left(\begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma \right), \quad \varepsilon_{ij} \sim N(0, \sigma^2)
\end{aligned}
$$

- We assess impact of alternative priors for Σ
  - using simulations
  - with a real data set

Problems with the conjugate option

Inverse Wishart distribution is conjugate and usually available in software.

However,

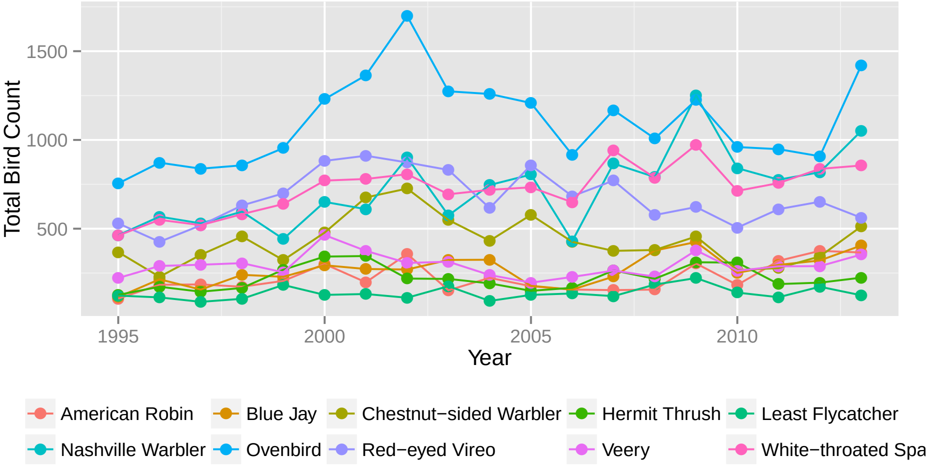
- "*implies the same amount of prior information about each of the variance parameters in the covariance matrix*" (Gelman et al., 2003).
- relation between ρ_{ij} and σ_i, higher values for the standard deviation σ_i are associated with higher correlations, ρ_{ij} close to 1 or -1 (Tokuda et al., 2011).

Also, we have found in a scenario with small variability,

- underestimate correlation
- overestimate standard deviation

Bird counts on Superior forests

The Natural Resources Research Institute (University of Minnesota Duluth) carry out monitoring program for study regional population trends of forest birds.

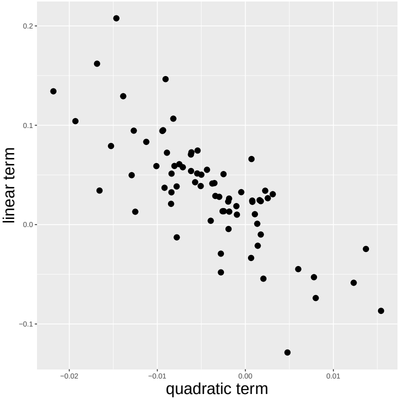


Quadratic trend model

- y_{st}: bird count for spacies s in year t.
- OLS regression model for each species

$$y_{st} \sim N(\beta_{0s} + \beta_{1s} t + \beta_{2t} t^2, \sigma^2)$$

- $\mathrm{Corr}(\hat{\beta}_{1s}, \hat{\beta}_{2s}) = -.77$



Quadratic trend model

Use a Bayesian hierarchical linear model with IW prior.

$$
\begin{aligned}
y_{st} &\sim N(\beta_{0s} + \beta_{1s} t + \beta_{2t} t^2, \sigma^2) \\
\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma \right) \\
\Sigma &\sim IW(d+1, I)
\end{aligned}
$$

Define $\rho = \Sigma_{23} / \sqrt{\Sigma_{22}\Sigma_{33}}$.

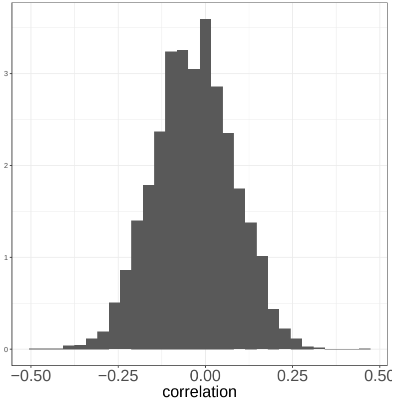
Quadratic trend model

Use a Bayesian hierarchical linear model with *IW* prior.

$$
\begin{aligned}
y_{st} &\sim N(\beta_{0s} + \beta_{1s} t + \beta_{2t} t^2, \sigma^2) \\
\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma \right) \\
\Sigma &\sim IW(d+1, I)
\end{aligned}
$$

Define $\rho = \Sigma_{23} / \sqrt{\Sigma_{22}\Sigma_{33}}$.

How does $p(\rho|y)$ look like?



Multivariate normal model

A simple model:
Consider n observations from $Y_i \sim N_d(0, \Sigma)$ distribution. Likelihood function can be written as follows:

$$p(y|\mu, \Sigma) \propto |\Sigma|^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n} y_i' \Sigma^{-1} y_i} = |\Sigma|^{-n/2} e^{-\frac{1}{2} tr(\Sigma^{-1} S_0)}$$

where $y_i \in R^d$ a realization of Y_i, $S_0 = \sum_{i=1}^{n} y_i y_i'$, individual entries of Σ are $\Sigma_{ii} = \sigma_i^2$ and $\Sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$.

We compare alternative covariance matrix priors in this context.

Alternative Σ priors

$IW(\nu, \Lambda)$ Inverse Wishart: Λ is matrix parameter related to location and ν degrees of freedom parameter.

$$p(\Sigma) \propto |\Sigma|^{-(\nu+d+1)/2} e^{-\frac{1}{2} tr(\Lambda \Sigma^{-1})}$$

Alternative Σ priors

$IW(\nu, \Lambda)$ Inverse Wishart: Λ is matrix parameter related to location and ν degrees of freedom parameter.

$$p(\Sigma) \propto |\Sigma|^{-(\nu+d+1)/2} e^{-\frac{1}{2} tr(\Lambda\Sigma^{-1})}$$

$SIW(\nu, \Lambda, b, \delta)$ Scaled inverse Wishart: $\Sigma = \Delta\, Q\, \Delta$ where $\Delta_{ii} = \xi_i$, then

$$Q \sim IW(\nu, \Lambda) \qquad \log(\xi_i) \overset{ind}{\sim} N(b_i, \delta_i)$$

Alternative Σ priors

$IW(\nu, \Lambda)$ Inverse Wishart: Λ is matrix parameter related to location and ν degrees of freedom parameter.

$$p(\Sigma) \propto |\Sigma|^{-(\nu+d+1)/2} e^{-\frac{1}{2} tr(\Lambda \Sigma^{-1})}$$

$SIW(\nu, \Lambda, b, \delta)$ Scaled inverse Wishart: $\Sigma = \Delta \, Q \, \Delta$ where $\Delta_{ii} = \xi_i$, then

$$Q \sim IW(\nu, \Lambda) \qquad \log(\xi_i) \stackrel{ind}{\sim} N(b_i, \delta_i)$$

$HIW_{ht}(\nu, \lambda, \delta)$ Hierarchical inverse Wishart: Λ diagonal matrix, $\Lambda_{ii} = \lambda_i$,

$$\Sigma | \lambda \sim IW(\nu + d - 1, 2\nu\Lambda) \quad \lambda_i \stackrel{ind}{\sim} \text{Ga}\left(\frac{1}{2}, \frac{1}{\delta_i^2}\right) \quad E(\lambda_i) = \frac{\delta_i^2}{2}$$

Alternative Σ priors

$IW(\nu, \Lambda)$ Inverse Wishart: Λ is matrix parameter related to location and ν degrees of freedom parameter.

$$p(\Sigma) \propto |\Sigma|^{-(\nu+d+1)/2} e^{-\frac{1}{2}tr(\Lambda\Sigma^{-1})}$$

$SIW(\nu, \Lambda, b, \delta)$ Scaled inverse Wishart: $\Sigma = \Delta\ Q\ \Delta$ where $\Delta_{ii} = \xi_i$, then

$$Q \sim IW(\nu, \Lambda) \qquad \log(\xi_i) \overset{ind}{\sim} N(b_i, \delta_i)$$

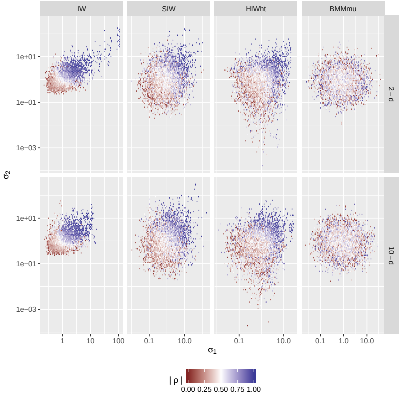
$HIW_{ht}(\nu, \lambda, \delta)$ Hierarchical inverse Wishart: Λ diagonal matrix, $\Lambda_{ii} = \lambda_i$,

$$\Sigma|\lambda \sim IW(\nu + d - 1, 2\nu\Lambda) \quad \lambda_i \overset{ind}{\sim} \text{Ga}\left(\frac{1}{2}, \frac{1}{\delta_i^2}\right) \quad E(\lambda_i) = \frac{\delta_i^2}{2}$$

$BMM_{mu}(\nu, \Lambda, b, \delta)$ Separation strategy: $\Sigma = \Lambda\ R\ \Lambda$ where $\Lambda_{ii} = \sigma_i$ and $R = \Delta^q Q \Delta^q$ with $\Delta^q_{ii} = Q_{ii}^{-1/2}$

$$Q \sim IW(\nu, I) \quad \log(\sigma_i) \overset{ind}{\sim} N(b_i, \delta_i)$$

Samples from prior



Positive relationship among σ_1 and σ_2, also large $|\rho_{12}|$ values appear when the two variances are high.

Impact on the posterior inference

We simulate normally distributed data,

$$Y \sim N_d(0, \Sigma)$$

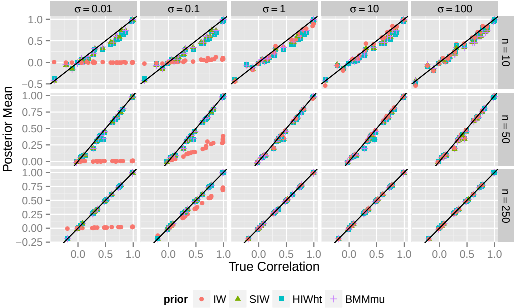
where d represent the dimension, we use $(\Sigma)_{ii} = \sigma \;\; \forall i \in \{1, \ldots, d\}$ and $(\Sigma)_{ij} = \sigma^2 \rho$ which implies all variances and all correlations are equal.

Table: Simulation scenarios. Specific values used in simulations for each parameter.

	Bivariate	Ten-dimensional
Sample size (n)	10,50,250	10,50
Standard deviation (σ)	0.01, 0.1, 1, 10, 100	0.01, 1, 100
Correlation (ρ)	0, 0,25, 0.5, 0.75, 0.99	0, 0.99

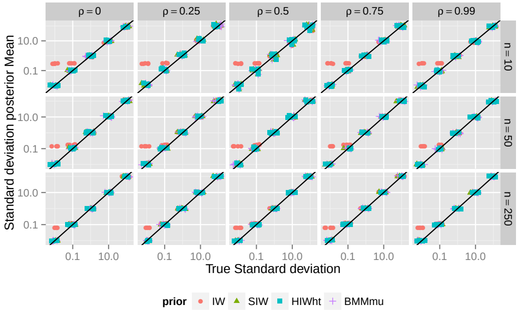
Each scenario is replicated 5 times.

Inference for ρ_{12}



When standard deviation is small, $\sigma = 0.01$ or $\sigma = 0.1$ the IW prior heavily shrinks the posterior correlation towards 0 even if the true correlation is close to 1.

Inference for σ_1



IW prior overestimate the standard deviation when its true value is very low

Summary

1. IW prior is restrictive,
   - correlations are small when variances are small
   - variances are positive correlated
   - Posterior inference with IW may be biased (in low variance case)
2. SIW and HIW_{ht} shows similar characteristics, but more flexible.
3. BMM_{mu} is the most flexible, variances and correlations are independent.
4. Posterior inference with BMM_{mu}, SIW or HIW_{ht} is fine.

Discussion

Prior choice

1. When it is possible to use a HMC sampler BMM_{mu} proposed by Barnard et al. (2000) gives modeling flexibility and good inferences properties.
2. Whenever we use Gibbs base samplers (as JAGS or BUGS) a prior which maintain conjugacy might be preferable such as the scaled inverse Wishart.
3. If we are constraint to use IW, we may recommend to scale the data first in order to avoid possible biased estimates for correlations.

Discussion

Prior choice

1. When it is possible to use a HMC sampler BMM_{mu} proposed by Barnard et al. (2000) gives modeling flexibility and good inferences properties.
2. Whenever we use Gibbs base samplers (as JAGS or BUGS) a prior which maintain conjugacy might be preferable such as the scaled inverse Wishart.
3. If we are constraint to use IW, we may recommend to scale the data first in order to avoid possible biased estimates for correlations.

Future steps

Different Model Hierarchical linear model context.

Different Priors Use LKJ prior for the correlation matrix.
Use other distributions for IW parameters ν and Λ.

References

Barnard, J., McCulloch, R., and Meng, X.-L. (2000), "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage," *Statistica Sinica*, 10, 1281–1312.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian data analysis*, Chapman and Hall.

Tokuda, T., Goodrich, B., Van Mechelen, I., and Gelman, A. (2011), "Visualizing Distributions of Covariance Matrices," http://www.stat.columbia.edu/~gelman/research/unpublished/Visualization.pdf.