Problems of the Inverse Wishart

immediate

Unknown (?)

1 Introduction

- Σ priors comonlly used
- IW problems: lack of flexibility becasuse 1 df parameter, prior dependency among correlation and variances. Also the low density region for inverse-chi2 mention in Gelman.
- Focus: impact on posterior inferences of these problems, specially the last one.
- IW "incidence": R packages or articles that use this prior.

The natural conjugate prior for normally distributed data is the Inverse-Wishart distribution. There are two main problems that been pointed out with this prior. First the uncertainty of all the variances is controlled by one single degrees-of-freedom parameter, which may results in lacks of flexibility for the inference (see Sec. 19.2 of Gelman et al. (2003)). Secondlly, it impose a prior dependence amnog variances and correlations (Tokuda et al. 2011).

Gelman (2006) shows scaled inv-χ^2 distribution has extremely low density in a region near zero and thus causes bias in the result posteriors for these variances. This issue is extended to the inverse wishart case since the implied distribution on each individual variance is actually inv-χ^2.

These issues are focus on characteristics of the prior. However, the impact of these characteristics on the posterior distribution and the inference are not studied that much. We show how the inference about individual vairances and correlations is afected when we use IW as a prior. The main impact is due to the low density region of the inv-χ^2 density which may explain bias in the posterior inference for Σ when the true variances are low.

Although all the problems that IW prior has, is still widelly used in aplications. Several R packages has built in functions that uses IW as a prior where the user only may be able to change the hyperparameter values[1] (and sometimes not even this). In this sense it would be nice to find some solution to its problems as a prior without designing a new one.

2 Statistical model

- Describe the likelyhood and the prior
- Obtain posterior for Σ, σ_i^2 and ρ_{ij}
- Study LDR on IG density (via inverse function or cdf)
- study hypergeometric function (appears on ρ posterior)
- Show posterior bias when vairance is low (simulations ?). Relate the bias with characteristic in the prior, lwd and hypergeometric function.

Consider the multivariate normal model, that is let $Y_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$ and assume $Y_i \overset{iid}{\sim} N(\mu, \Sigma)$ with $\mu \in \mathbb{R}^d$ and Σ is a d-dimensional positive definite matrix.

$$p(y|\mu,\Sigma) \propto |\Sigma|^{-n/2}\text{exp}\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^\top\Sigma^{-1}(y_i-\mu)\right\} = |\Sigma|^{-n/2}\text{exp}\left\{-\frac{1}{2}\text{tr}(\Sigma^{-1}S_\mu)\right\} \quad (1)$$

The likelihood is provided in equation (1) where y represents the entire data and $S_\mu = \sum_{i=1}^n (y_i-\mu)(y_i-\mu)^\top$. Which can be decompose as $S_\mu = A + M = \sum_{i=1}^n (y_i-\bar{y})(y_i-\bar{y})' + n(\bar{y}-\mu)(\bar{y}-\mu)^\top$.

2.1 Posterior density

The primary parameter of interest is the matrix Σ with elements Σ_{ij}. We will often refer to the standard deviations σ_i and correlations ρ_{ij} where $\sigma_i^2 = \Sigma_{ii}$ and $\Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$.

To complete model we set uniform prior for μ, $p(\mu) \propto 1$ and inverse wishart prior on $\Sigma \sim IW(\nu, \Lambda)$, i.e. $p(\Sigma) \propto |\Sigma|^{-(\frac{\nu+d+1}{2})}e^{-\frac{1}{2}tr(\Lambda\Sigma^{-1})}$. Posterior distribution for Σ is derived as

[1]packages: `nicheROVER,SharedHT2 (not in cran anymore), Boom, sbgcop, MCMCglmm,BayesComm, MSBVAR,bayesSurv,phcfM, monomvn, miscF, bayesm, spBayes, factorQR, agRee, DPpackage`

follows,

$$
\begin{aligned}
p(\Sigma|y) \quad & \propto \int p(\Sigma,\mu|y)d\mu \\
& \propto p(\Sigma)\int |\Sigma|^{-n/2}e^{-\frac{1}{2}tr(\Sigma^{-1}(A+M))}d\mu \\
& \propto p(\Sigma)|\Sigma|^{-n/2}e^{-\frac{1}{2}tr(\Sigma^{-1}A)}\int e^{-\frac{1}{2}tr(\Sigma^{-1}M)}d\mu \\
& \propto p(\Sigma)|\Sigma|^{(-n/2+1/2)}e^{-\frac{1}{2}tr(\Sigma^{-1}A)} \\
& \propto |\Sigma|^{-(\frac{\nu+d+n}{2})}e^{-\frac{1}{2}tr(\Sigma^{-1}(A+\Lambda)}
\end{aligned}
\tag{2}
$$

implying that $\Sigma|y \sim IW(n+\nu-1, A+\Lambda)$ as expected.

The posterior ditribution for individual vairances is directlly obtained from IW properties as a scaled inverse chi-square distribution[2] for each variance $\sigma_i^2 \sim \text{inv-}\chi^2(\nu-d+1, \frac{\lambda_{ii}}{\nu-d+1})$ where λ_{ii} is the i^{th} diagonal entry of Λ.

Another IW property is that any submatrix is also distributed as IW, this fact can be used to obtain the posterior distribution of the correlation coeficient. In particular taking the first submatrix of dimension $d=2$ and letting $\tilde{A} = A+\Lambda$ we have

$$
p(\sigma_1^2,\sigma_2^2,\sigma_{12}) \propto (\sigma_1^2\sigma_2^2-\sigma_{12}^2)^{-\frac{n+2}{2}} exp\left[-\frac{1}{2(\sigma_1^2\sigma_2^2-\sigma_{12}^2)}(\tilde{a}_{11}\sigma_2^2 - 2\tilde{a}_{12}\sigma_{12} + \tilde{a}_{22}\sigma_1^2\right]
$$

where $a_{kj} = \sum_{i=1}^{n}(y_{ik}-\bar{y}_k)(y_{ij}-\bar{y}_j)$ and $\tilde{a}_{ij} = a_{ij}+\lambda_{ij}$.

Then we follow the derivations in Box and Tiao (1973) to obtain the posterior density of the correlation when a Jefrey prior is consider. They used a transformation proposed by Fisher to obtain the sampling distribution of the pearson correlation coeficient. Let $r = a_{12}/\sqrt{a_{11}a22}$ and $r = \tilde{a}_{12}/\sqrt{\tilde{a}_{11}\tilde{a}22}$

$$
x = (\tfrac{\sigma_1\sigma_2}{\tilde{a}_{11}\tilde{a}_{22}})^{1/2} \quad w = (\tfrac{\sigma_1\tilde{a}_{22}}{\tilde{a}_{11}\sigma_2})^{1/2} \quad \rho = \tfrac{\sigma_{12}}{\sigma_1\sigma_2} \quad J = \tfrac{2x^2(\tilde{a}_{11}\tilde{a}_{22})^{3/2}}{w}
$$

$$
\begin{aligned}
p(x,w,\rho) \quad & \propto (x^2(1-\rho^2))^{-\frac{n+2}{2}}\tfrac{x^2}{w}exp\left[-\tfrac{1}{2x(1-\rho^2)}(w^{-1}-2\tilde{r}\rho w)\right] \\
p(w,\rho) \quad & \propto \tfrac{(1-\rho^2)^{-\frac{n+2}{2}}}{w}\int x^{-n}exp\left[-\tfrac{1}{2x(1-\rho^2)}(w^{-1}-2\tilde{r}\rho w)\right]dx \\
p(w,\rho) \quad & \propto \tfrac{(1-\rho^2)^{-\frac{n+2}{2}}}{w}(1-\rho^2)^{n-1}(w^{-1}-2\tilde{r}\rho w)^{-(n-1)}
\end{aligned}
$$

Finnally,

$$
p(\rho|y) \propto (1-\rho^2)^{\frac{n+\nu}{2}-2}\int_0^\infty w^{-1}(w^{-1}-2\tilde{r}\rho+w)^{-(n+\nu-1)}dw \tag{3}
$$

[2]The scaled inverse chi-square denoted by $X \sim \text{inv-}\chi^2(\nu,s^2)$ has a density function given by $p(x) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}s^\nu x^{-(\nu/2+1)}\exp\left\{-\nu s^2/2x\right\}$

2.2 useful matemathical functions

lambert W and hypergeometric F ...

2.3 Understanding the inference bias

show why posterior is biased when true variance is low...

3 Solutions ...

- uses small values for diagonal elements in the prior, instead of I matrix.
- use sample vairance on the pior matrix.

References

Box, G. E. and Tiao, G. C. (1973), "Bayesian Inference in Statistical Analysis," Tech. rep., DTIC Document.

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, 1, 515–533.

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`.