Searching for Ideal Priors for Covariance Matrices

Mohsen Pourahmadi
Division of Statistics
Northern Illinois University

SAMSI Bayesian Focus Week

Oct. 30 - Nov. 3, 2006

Outline

1. Unconstrained Parameterization and Priors: GLM

2. ARMA Models & Stationary Processes:

   Toeplitz matrices & partial correlations

3. Variance-Correlation Decomposition: Partial correlations

4. Spectral Decomposition: Orthogonal matrices

5. Cholesky Decompositions: A sequence of regressions

6. Conclusions

1. Unconstrained Parameterization & Priors: GLM

- The theory of generalized linear models (GLM) underscores the importance of introducing **unconstrained** and **interpretable** parameters. The link function is usually not unique, it acts *componentwise* on the mean vector μ.

- The revolution in Bayesian computation in the early 1990's has encouraged going beyond the standard **conjugate priors** (inverse Wishart) for covariance matrices.

- Various decompositions of covariance matrices lead to flexible priors and partially unconstrained reparameterizations. Some of these are **unique** up to a **rotation** as in factor analysis. The most interpretable ones rely on the concepts of partial correlation/regressions.

2. **ARMA Models & Stationary Processes:**

- $AR(2): \ y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t.$

Stationary Region: All $(\phi_1, \phi_2) \epsilon R^2$ so that the roots of

$$1 - \phi_1 \ z - \phi_2 \ z^2 = 0,$$

lie outside the unit circle.

$S_2 : -1 < \phi_2 < 1 \ , \ -1 + \phi_2 \ < \phi_1 < 1 - \phi_2.$

Since $-1 < \phi_2 < 1 \ , \ -1 < \frac{\phi_1}{1-\phi_2} < 1,$
Using the Fisher z-transform of ϕ_2 and $\dfrac{\phi_1}{1-\phi_2}$ gives the unconstrained region

$$S_2^* : \ \phi_2^* = \log \frac{1-\phi_2}{1+\phi_2} \ \epsilon R, \ \phi_1^* = \log \frac{1-\phi_2+\phi_1}{1-\phi_2-\phi_1} \epsilon R.$$

Now, normal priors can be introduced for $(\phi_1^*, \ \phi_2^*)$ as in Marriot & Smith (1992). For more general polynomials, use the fact that any polynomial can be factored in terms of polynomials of degree ≤ 2.

• Stationary Processes & Partial Correl.

- Durbin-Levinson Algorithm (Levinson, 1947, Durbin, 1960) connects the covariance function $\{\gamma_k\}$ of a stationary process to its partial correlation function $\{\pi_k\}$ in a natural way. The latter quantities vary freely in the range $[-1, 1]$.

- R. H. Jones (1980) used partial correlations to reparameterize ARMA models. It has been the backbone of most ARMA fitting procedures in software packages since 1980.

- F. L. Ramsey (1974):

  There is a one-to-one correspondence between the partial **correlation function** $\{\pi_k\}$ and the **covariance function** $\{\gamma_k\}$ of a stationary process.

3. Variance-Correlation Decomposition: $\Sigma \leftrightarrow (R, D)$

$$\Sigma = D\ R\ D$$
$$D = diag\left(\sqrt{\sigma_{11}}, \cdots, \sqrt{\sigma_{pp}}\right)$$
$$R = (\rho_{ij}), \rho_{ij} = corr(y_i, y_j).$$

- $\log \sqrt{\sigma_{ii}}$ is unconstrained, but the entries of R are constrained.

- How to assign priors to or simulate R?

  The common practice is to assign priors to R or to each of its entries and simulate (draw) one entry at a time with a view to keep R positive definite.

• Barnard, McCulloch & Meng (2000):

- $p(\Sigma) = p(R, D) = p(R|D)p(D)$.
  i. R & D can be assumed independent,
  ii. the dist. of R is exchangeable or invariant to permuation of the indices,
  iii. use diffuse priors on R.

- Two suggested choices for $p(R)$ are
  i. a marginally uniform prior,
  ii. the jointly uniform prior
- A suggested alternative:
  uniform priors on the partial correlations

- They use Gibbs sampler and draw a particular correlation $\rho_{ij} = \rho$, given the other entries of R.

  Q. Starting with R pd, what values of $\rho_{ij} = \rho$ will keep $R(\rho)$ positive definite?

  A. (1) $R(\rho)$ is pd iff $|R(\rho)| > 0$.

  (2) $|R(\rho)| = a\rho^2 + b\rho + c$,

  where a, b and c do not depend on ρ, and

$$c = |R(0)| \ , \quad b = \frac{|R(1)| \ - \ |R(-1)|}{2},$$

$$a = \frac{|R(1)| + |R(-1)| \ - 2|R(0)|}{2}$$

  (3) $a\rho^2 + b\rho + c > 0$ gives the interval of values of ρ which keeps $R(\rho)$ pd.

- This requires computing many determinants, is not fast when p is large. Wong, Carter and Kohn (2003) give a faster method of computing a,b,c using the Cholesky decomposition of R.

- Wong, Kohn & Carter (2003): $\Sigma^{-1} = DRD \leftrightarrow (R, D)$

- For D and R from Σ^{-1}, their nonredundant entries are the **partial variances** $var(y_i|y_k,\ k \neq i)$,
  and the **partial correlations** $corr(y_i, y_j|y_k, k \neq i, j)$.

- Covariance selection (Dempster, 1972):
  i. Set certain off-diagonal entries of Σ^{-1} to zero.
  ii. Random effect selection requires certain diagonal entries to be zero.
- Gaussian graphical models

- Covariance selection priors (Kohn et al. 2003, 2006)

  Make Σ^{-1} sparse by forcing certain entries of R to be zero. Works for decomposable (Giudici & Green, 1999) and nondecomposable (Roverato, 2002, Atay-Kayis & Massam, 2006, $\cdots$) graphs.

  Liechty, Liechty and Muller (2004); Liu & Daniels (2006).

- "Sequential" Partial Correlations

- H. Joe (2006) Reparametrizes R in terms of

$$\rho_{i,i+1} \ , \quad i = 1, \cdots, p-1,$$

and the partial correlations

$$\rho_{i,j|i+1,\cdots,j-1} \ = \ \rho_{i,j|int(i,j)} \ , \quad i - j \geq 2,$$

where these vary freely in $(-1, 1)$.

Theorem . $|R| = \prod_{k=1}^{p-1} \prod_{i=1}^{p-k} \left(1 - \rho^2_{i,i+k}|_{int(i,i+k)}\right)$.

Examples .

For $p = 3$, $|R| = \left(1 - \rho^2_{12}\right)\left(1 - \rho^2_{23}\right)\left(1 - \rho^2_{13|2}\right)$.

For $p = 4$, $|R| = \left(1 - \rho^2_{12}\right)\left(1 - \rho^2_{23}\right)\left(1 - \rho^2_{34}\right)$

$$\left(1 - \rho^2_{13|2}\right)\left(1 - \rho^2_{24|3}\right)\left(1 - \rho^2_{14|23}\right).$$

- A prior for R or a random pd correlation matrix can be generated by choosing independent distributions $g_{ij}(\cdot)$ on $(-1, 1)$ for these new parameters. More precisely,

$$f\left(\rho_{ij,1\leq i<j\leq p}\right) = \prod_{k=1}^{p-1}\prod_{i=1}^{p-k} g_{i,i+k}(\rho_{i,i+k|int(i,i+k)})\times|J_p|,$$

where $|J_p|$ is the determinant of the Jacobian for the transformation from $(\rho_{ij}, 1 \leq i < j \leq p)$ to $(\rho_{ij|int(i,j)})$.

- It turns out that

$$|J_p| = \left[\prod_{k=1}^{p-2}\prod_{i=1}^{p-k}\left(1-\rho^2_{i,i+k|int(i,i+k)}\right)^{p-1-k}\right]^{-1/2}.$$

Note that the exponents *depend on k, but not on i.* In a sense, **stationarity** is implied along the kth diagonal.

- This suggests to choose $g_{i,i+k}(\cdot)$ as certain beta density to achieve a simple form for f, the joint density of $R = (\rho_{ij})$. Recall the linearly transformed symmetric Beta (α, α) on $(-1, 1)$ with the density:

$$g(u) = 2^{-2\alpha+1}[B(\alpha,\alpha)]^{-1}\left(1-u^2\right)^{\alpha-1}, |u| < 1, \alpha > 0.$$

- Choosing $B(\alpha_k, \alpha_k)$ for $\rho_{i,i+k|int(i,i+k)}$, it follows that f is proportional to

$$\prod_{k=1}^{p-1} \prod_{i=1}^{p-k} \left(1 - \rho^2_{i,i+k|int(i,i+k)}\right)^{\alpha_{k-1}-(p-1-k)/2}.$$

Theorem . With $\alpha_k = \alpha + \frac{1}{2}(p-1-k)$,

(a) the above density is proportional to

$$\prod_{k=1}^{p-1} \prod_{i=1}^{p-k} \left(1 - \rho^2_{i,i+k|int(i,i+k)}\right)^{\alpha-1} = |R|^{\alpha-1}.$$

(b) the same density would arise if the indices of the correlation matrix were permuted before computing the partial correlations along the kth diagonal, $k = 1, \cdots, p-1$.

- Partial Correlation Vines:

- Kurowicka and Cooke (2006). Completion problem with partial correlation vines. Linear Algebra and its Applications, 418, 188-200.

- Theorem . There is a one-to-one correspondence between the set of $p \times p$ positive-definite **correlation matrices** and the set of **partial correlations for any regular vine** on p elements.

- A regular vine is a tool for picking out those partial correlations which uniquely determine the correlation matrix.

4. The Spectral Decomposition:
   $\sum = P\Lambda P'$,
   where P orthogonal matrix , $\Lambda = diag(\lambda_1, \cdots, \lambda_p)$,

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p$$

   , are the **distinct eigenvalues** of $\sum$ and the columns of P are the respective normalized eigenvectors.

- $\Sigma \leftrightarrow (P, \Lambda)$

- Of course, P is constrained by orthogonality.

Q. How to make it unconstrained or assign priors to its entries?

A. (1) **Givens angles:** Yang and Berger (1994); reference priors
Daniels and Kass (1999); normal priors on the logit.

(2) Householder reflections: Pinheiro (1994); Bates & Pinheiro (1996), closely related to the Givens angles

- These are hard to interpret statistically.

(3) Local Parametrization : Boik (1998, 2002)

- Hard to explain in a short time.

(4) Log Parametrization : A matrix P is **orthogonal**, iff $P = e^U$ where U is **skew-symmetric** (Muirhead, 1982).

- Thus, $\log P = U$, has **unconstrained** parameters, hard to interpret, has not been used much.

- Logarithmic Models for Σ (Leonard et al. 1992, 1996)

- A symmetric matrix Σ is pd iff $\Sigma = e^A$,
  where A is a real symmetric matrix with **unconstrained** entries.

- Entries of A are hard to interpret.

  Brown, Zidek and Li (1994); Yang and Berger (1994)

  Pinheiro and Bates (1996).

- Finally, the reparametrization of Σ in terms of (P, Λ) is reasonable so long as the eigenvalues are **distinct**. The situation is not that clear, otherwise.

  Example:

(i.) $\Sigma = \sigma^2 I$, $\Sigma = (P, \sigma^2 I)$, for **any** orthogonal matrix P.

(ii.) Compound symmetry, $\lambda_1 = 1 + (p-1)\rho$,

$$\lambda_2 = \cdots = \lambda_p = 1 - \rho.$$

(iii.) Spiked covariances.

5. Cholesky Decompositions:

In most textbooks and sofware packages, the Cholesky decomposition of $\sum$ is of the form

$$\sum = CC',$$

where C is a unique lower triangular matrix with positive diagonal entries (Cholesky, 1918; Bartlett, 1933). It has been used frequently in optimization problems involving *pd* matrices (Kalman, 1961; Lindstrom and Bates, 1988, 1990: linear and nonlinear mixed models, $\cdots$).

Interpretation of the entries of $C = (c_{ij})$ is difficult (Bates and Pinheiro, 1996). However, reducing C to unit lower triangular matrices through post(pre)-multiplication by the inverse of

$$D = diag(c_{11}, \cdots, c_{pp}),$$

makes the task of interpretation much easier (Pourahmadi, 1999; Chen & Dunson, 2003).

$$\sum = \underbrace{CD^{-1}}\ DD\ \underbrace{D^{-1}C'} = D\ \underbrace{D^{-1}C}\ \underbrace{C'D^{-1}}\ D.$$

- Priors for $\sum = (L, D)$ or $(\tilde{L}, D)$

- Brown, Zidek and Li (1994): Sequential Priors

Multivariate normal / Inverse Wishart.

- Pourahmadi & Daniels (2002): Sequential Priors

Normal / Inverse gamma .

- Smith & Kohn (2002): Variable selection priors

- Chen & Dunson (2003): Random effect selection priors on the diagonal entries of D. In the context of linear mixed models (LMM).

- Daniels & Zhao (2003): Normal / Inverse gamma in LMM.

- Cai & Dunson (2006): Priors on (L, D) in GLMM.

- Dorba & West (2004): Variable selection priors on L in the context of graphical models.

- Huang et al. (2006): Penalized normal likelihood with LASSO penalty (Laplace priors on L).

- **Time Series & Cholesky Decomposition:**
  The AR(2) model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t,$$

  for $t = 1, 2, \ldots, n$ can be written as

$$\begin{array}{rcrcrcl} y_1 & = & \phi_1 y_0 & + & \phi_2 y_{-1} & + & \varepsilon_1, \\ y_2 & - & \phi_1 y_1 & = & \phi_2 y_0 & + & \varepsilon_2, \\ \vdots & & & & & & \vdots \\ y_p & - & \phi_1 y_{p-1} & - & \phi_2 y_{p-2} & = & \varepsilon_p. \end{array}$$

  Setting $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)$ and $e = (y_{-1}, y_0)$, it becomes the **regression-like model**

$$TY = \varepsilon + Ke,$$

  where

$$T = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ -\phi_1 & 1 & 0 & \cdots & \cdots & 0 \\ -\phi_2 & -\phi_1 & 1 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\phi_2 & -\phi_1 & 1 \end{bmatrix}, \; K = \begin{bmatrix} \phi_2 & \phi_1 \\ 0 & \phi_2 \\ \cdots & \cdots \\ 0 & \cdots \; 0 \\ \vdots & \vdots \\ 0 & \cdots \; 0 \end{bmatrix} = \begin{bmatrix} K_1 \\ \cdots \\ 0 \end{bmatrix}.$$

  When ε and e are independent, it follows that

$$\begin{aligned} T\mathrm{cov}(Y)T' &= \sigma^2 I_p + \begin{pmatrix} K_1 \mathrm{cov}(e) K_1' & 0 \\ 0 & 0 \end{pmatrix} \\ &= \text{A nearly } \textbf{diagonal} \text{ matrix.} \end{aligned}$$

- Now, we force K to be zero using the idea of regression.

- **Reg./G.-Schmidt/Chol./Szegö/Bartlett/DL/KF**
  Regress y_t on its predecessors:

$$y_t = \phi_{t,t-1}y_{t-1} + \cdots + \phi_{t1}y_1 + \varepsilon_t,$$

y_1	y_2	y_3	$\cdots$	y_{p-1}	y_p
σ_1^2					
ϕ_{21}	σ_2^2				
ϕ_{31}	ϕ_{32}	σ_3^2			
$\vdots$	$\vdots$		$\ddots$		
ϕ_{p1}	ϕ_{p2}	$\cdots$	$\cdots$	$\phi_{p,p-1}$	σ_p^2

  in matrix form

$$\begin{bmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{p1} & -\phi_{p2} & \cdots & -\phi_{p,p-1} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}, Ty = \varepsilon,$$

$$T \Sigma T' = D,$$

$$\Sigma = T^{-1}DT'^{-1} = LDL'.$$

- This idea reduces the unintuitive task of modeling covariance to that of a sequence of regressions (with varying-order and varying-coefficients). Pourahmadi (1999, 2000) and Chen and Dunson (2003).

6. Conclusions

- Close connection between unconstrained parametrizations of covariance matrices and introducing flexible priors.

- Reparametrization in terms of partial correlations and regression coefficients.

- **Nonuniqueness** involving "**ordering**" of the variables in computing partial correlations, Cholesky decompositions, and spectral decomposition when the **eigenvalues are not distinct**.

- Can one take advantage of this nonuniqueness as in factor analysis, $\cdots$?