

Priors and models for correlation matrices

Mike Daniels, Mohsen Pourahmadi, Yanpin Wang

University of Florida and Texas A& M

June 2009

- 1 Background
- 2 Partial autocorrelations
- 3 Parsimonious modeling
- 4 Priors
- 5 Computations
- 6 Ongoing work

Problem

- reduce the dimension of a *correlation* matrix for longitudinal/ordered data
 - most practical when convenient parameterization and interpretable parameters
 - covariance matrix reparameterizations
 - Modified Choleski (GARP/IV)
 - log matrix/spectral decomp
 - problem: diagonal elements fixed (at one)
- specify appropriate priors for correlation matrix for longitudinal/ordered data
 - typically, correlations non-increasing with lag
- even if only mean/location of interest, dependence matters when incomplete data (see, e.g., Daniels and Hogan, 2008, Ch. 5)

Modified Choleski decomposition I

- Modified Choleski (GARP/IV) decomposition (Pourahmadi, 1999; Daniels and Pourahmadi, 2002): a parameterization popular for covariance matrices due to
 - computational simplicity
 - interpretation of the elements
- elements are the autoregressive coefficients and prediction variances from the sequential conditional distributions, $p(y_j | y_1, \dots, y_{j-1})$;
 - generalized autoregressive parameters (GARP) are the regression coefficient, ϕ in

$$E[Y_j | y_1, \dots, y_{j-1}] = \mu_j + \sum_{k=1}^{j-1} \phi_{jk}(y_k - \mu_k)$$

Modified Choleski decomposition II

- innovation (prediction) variances,

$$\text{Var}[Y_j \mid y_1, \dots, y_{j-1}] = \sigma_j.$$

- this parameterization does NOT work well for a correlation matrix since the unconstrained elements (for a covariance matrix), ϕ and $\log(\sigma_j)$ become highly constrained due to the marginal variances being fixed at one in a correlation matrix
- so, this parameterization loses much of its attraction
- is there a related alternative for a correlation matrix?

Partial autocorrelations

- parameterize a correlation matrix $R = \{\rho_{ij}\}$ in terms of the lag-1 correlations

$$\pi_{i,i+1} \equiv \rho_{i,i+1}, i = 1, \dots, p - 1$$

and the partial autocorrelations

$$\pi_{ij} = \rho_{ij|i+1, \dots, j-1} : j - i \geq 2$$

- replace the constrained matrix R by the simpler matrix $\Pi = \{\pi_{ij}\}$ with ones on the diagonal
- π_{ij} 's can vary freely in the interval $(-1, 1)$
- related work in Kurowicka and Cooke (2003, 2006) and Joe (2006)

Connection to marginal correlations I

Let $\mathbf{R}[j : j + k]$ be the submatrix of the correlation matrix \mathbf{R} containing rows and columns from j to $j + k$, i.e.

$$\mathbf{R}[j : j + k] = \begin{pmatrix} 1 & r'_1(j, k) & \rho_{jj+k} \\ r_1(j, k) & \mathbf{R}_2(j, k) & r_3(j, k) \\ \rho_{j+k,j} & r'_3(j, k) & 1 \end{pmatrix},$$

where

$$r'_1(j, k) = (\rho_{j,j+1}, \dots, \rho_{j,j+k-1}),$$

$$r'_3(j, k) = (\rho_{j+k,j+1}, \dots, \rho_{j+k,j+k-1}),$$

and $\mathbf{R}_2(j, k)$ is the correlation matrix corresponding to components $(j + 1, \dots, j + k - 1)$

Connection to marginal correlations II

Then, the partial autocorrelations between Y_j and Y_{j+k} adjusted for the intervening variables ($\pi_{j,j+k} = g(\rho_{j,j+k}, R[j:j+k])$)

$$\pi_{j,j+k} = \frac{\rho_{j,j+k} - r'_1(j, k)R_2(j, k)^{-1}r_3(j, k)}{D_{jk}} \quad (1)$$

where

$$D_{jk} = [1 - r'_1(j, k)R_2(j, k)^{-1}r_1(j, k)]^{1/2} * [1 - r'_3(j, k)R_2(j, k)^{-1}r_3(j, k)]^{1/2}$$

D_{jk} is a function of the partial variances.

Connection to marginal correlations III

- the function $g(\cdot)$ on the previous page that maps a correlation matrix R into the partial autocorrelation matrix Π , is invertible,
- solving for $\rho_{j,j+k}$

$$\rho_{j,j+k} = r_1'(j, k)R_2(j, k)^{-1}r_3(j, k) + D_{jk} \pi_{j,j+k}, \quad (2)$$

- easy to move between the two parameterizations

An Attractive Property

- the lag k partial autocorrelations are based on conditional regressions where the conditioning sets have the same number of elements
- i.e., they are *exchangeable* in some sense (all conditioning on $k - 1$ intervening variables)
- unlike similar sequential partial autocorrelations (regression coefficients) from Modified Choleski, $Y_j \mid y_1, \dots, y_{j-1}$
- simplifies building models for the partial autocorrelations, $\pi_{j,j+k}$ as a function of lag k and reducing the dimension in general

Parsimonious modeling

- transform partial autocorrelations to the entire real line using Fisher's z transform,

$$z(\pi_{jk}) = -\frac{1}{2} \log \frac{1 - \pi_{jk}}{1 + \pi_{jk}}$$

- avoids working with the complex constraints on
 - the correlation matrix R , directly parameterized with $\rho_{jk} = \text{corr}(Y_j, Y_k)$ (Barnard et al., 2000; Daniels and Normand, 2006)
 - or the matrix of full partial correlations constructed from Σ^{-1} (Wong et al. 2003), $\text{corr}(Y_j, Y_k | \{Y_l : l \neq j, l \neq k\})$

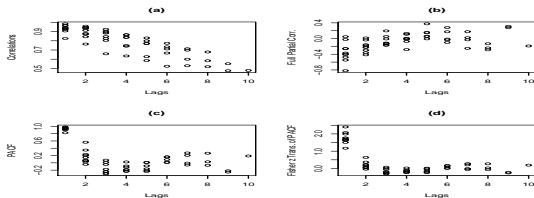
Parsimonious modeling

- $\pi_{j,j+k}$ gauges the conditional (on the intervening variables) dependence between variables k units apart
- expect them to be smaller for larger k .
- also expect a lot of partial autocorrelations to be zero
- e.g., AR(1) all partial autocor of lag 2 or higher are zero
- in the Bayesian framework, suggests putting shrinkage priors on the partial autocorrelations that shrink the matrix Π toward certain simpler structures (Daniels and Kass, 2001)
- in general, regression models

$$z(\pi_{jk}) = \mathbf{w}_{jk}\gamma. \quad (3)$$

where w_{jk} can be $w_{i,jk}$ (unit-specific covariates)

Exploratory plots for cattle data (Kenward)



For exploratory examine plots of

- $\{\pi_{j,j+k}; j = 1, \dots, p - k\}$ versus lag $k = 1, \dots, p - 1$
- $\{\pi_{j,j+k}; k = 1, \dots, p - 1\}$ vs. j for a fixed lag k (nonstationary)

Model and literature on priors I

Model:

$$Y_i \sim N_p(\mathbf{X}_i\beta, \mathbf{D}\mathbf{R}\mathbf{D})$$

where D is the diagonal matrix of standard deviations.

- Eaves and Chang (1992); reference prior for certain partial correlations
- Chib and Greenberg (1998); truncated multivariate normal distribution on the marginal correlations.
- Barnard et al. (2000); joint and marginal uniform
- Wong et al. (2003): prior on full partial correlations (more later), $\text{corr}(Y_j, Y_k | \{Y_l : l \neq j, l \neq k\})$
- Liechty et al. (2004): hierarchical priors; clustering of marginal correlations

Model and literature on priors II

- Zhang and Boscardin (2006); marginalized informative priors based on Wishart

Default priors

- Barnard Beta (joint uniform)
- Independent uniform
- Jeffreys'
- Marginal uniform

Barnard Beta prior I

$$p(R) \propto I\{|R| > 0, r_{ii} = 1, |r_{ij}| \leq 1\}$$

(Barnard et al, 2000)

- corresponds to the following (stationary) Beta priors

$$\pi_{i,i+k} \sim \text{Beta}_{[-1,1]}(\alpha_k, \alpha_k), \quad (4)$$

where $\alpha_k = 1 + \frac{1}{2}(p - 1 - k)$; see Joe (2006).

- shrinkage goes the opposite way of what want - more peaked for lag 1 and for lags greater than 1

Barnard Beta prior II

- such a prior results in the marginal priors for each of the marginal correlations being somewhat peaked around zero (same peakedness for *all* ρ_{jk}) - treats all the marginal correlations exchangeably
- Note that the priors become more peaked as p grows (α_k increases linearly in p ; centered at zero, variance decreases in p^3)
- for $p = 10$, prior probability less than $1/1000$ that $\pi > .8$
- for $p = 5$, prior probability of about $1/100$ that $\pi > .8$
- these are too extreme for ordered data (with serial correlation)

Proposal: Uniform on Π

- to offset the reversed shrinkage of standard priors, might try a simple alternative
- Uniform on $p(p-1)/2$ -dimensional hypercube (i.e., uniform on the transformed space)
- corresponds to independent $Unif(-1, 1)$ priors on partial autocorrelations

$$\pi_{jk} \sim Unif(-1, 1) = \text{Beta}_{[-1,1]}(1, 1)$$

Proposal: Uniform on Π

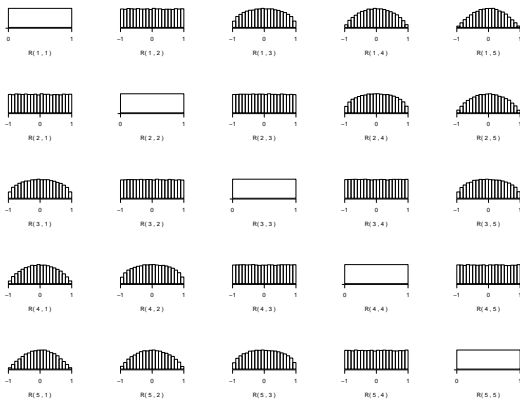
- induces the following prior for the correlation matrix R ,

$$p(R) \propto \left[\prod_{k=1}^{p-1} \prod_{i=1}^{p-k} (1 - \pi_{i,i+k}^2)^{p-1-k} \right]^{-1/2} \quad (5)$$

this is just the Jacobian from π to \mathbf{R}

- also induces an interesting behavior on the marginal correlations.

Marginal priors on ρ_{jk} from independent uniform priors on the partial correlations, π_{jk}



Independent Priors

- the priors on the (marginal) correlations ρ_{jk} , become gradually more peaked at zero as the lag $|j - k|$ grows.
- for $p = 15$, the (averaged) probabilities of being in $[-.5, .5]$, ordered from lag 1 to lag 14, are respectively,

(.50, .59, .65, .70, .73, .76, .78, .80, .82, .83, .84, .85, .86, .87)

- very desirable behavior for longitudinal data which typically exhibits serial correlation decaying with increasing lags.
- also appears that the priors for ρ_{jk} with $|j - k|$ fixed, appear to be the same (see the subdiagonals).

Stationary Priors I

Theorem

If the partial autocorrelations, π_{jk} have independent stationary priors, i.e

$$p(\pi_{jk}) = p(\pi_{il}) \text{ if } |j - k| = |i - l|, \quad (6)$$

(or the priors are the same along the subdiagonals of Π), then the marginal priors on the correlations ρ_{jk} are also stationary, i.e.

$$p(\rho_{jk}) = p(\rho_{il}) \text{ if } |j - k| = |i - l|. \quad (7)$$

(or the priors are the same along the corresponding subdiagonal of R).

Stationary Priors II

Theorem implies that independent 'stationary' priors on Π induce 'stationary' priors on the marginal correlations.

Simulations: description

- three true matrices

- 1 AR(1)[$\rho = .8$]

- 2 AR(1)[$\rho = .6$]

- 3 $\pi_{k,k+1} = .8, \pi_{k,k+2} = .4, \pi_{k,k+3} = .1, \pi_{k,k+j} = 0 : j > 3$
($\rho_{k,k+1} = .8, \rho_{k,k+2} = .78, \rho_{k,k+3} = .73, \rho_{k,k+4} = .68$)

- $p = 5$ ($n = 10, 25, 50, 100$); $p = 10$ ($n = 25, 50, 100$)

- losses

- LL: $tr(\hat{\mathbf{R}}\mathbf{R}^{-1}) - \log |\hat{\mathbf{R}}\mathbf{R}^{-1}| - p$

- SEL on Fisher's z-transform of partial autocorrelations

- SEL on Fisher's z-transform of marginal correlations

- priors

- Barnard Beta (BB)

- Independent uniform (IU)

Simulations: Results ($p = 5$)

R	LL		SEL-P		SEL-M	
	BB	IU	BB	IU	BB	IU
AR(.8)	2.4	1.6	1.4	1.1	1.7	1.1
	.69	.51	.48	.39	.55	.38
	.27	.23	.20	.18	.22	.17
	.11	.10	.09	.08	.09	.07
AR(.6)	1.1	.99	.89	.86	.98	.89
	.48	.45	.41	.39	.46	.44
	.23	.21	.19	.18	.19	.17
	.11	.11	.10	.09	.10	.10
Full	2.7	1.9	1.8	1.3	2.8	1.8
	.68	.50	.46	.35	.61	.38
	.25	.20	.19	.16	.21	.15
	.12	.11	.10	.09	.08	.07

Simulations: Results ($p = 10$)

R	LL		SEL-P		SEL-M	
	BB	IU	BB	IU	BB	IU
AR(.8)	5.0	2.8	3.3	2.3	5.2	2.7
	1.6	1.1	1.2	.88	1.7	.97
	.60	.48	.49	.42	.52	.36
AR(.6)	2.4	2.2	2.1	2.1	2.2	1.9
	.97	.85	.86	.79	.88	.71
	.49	.45	.45	.42	.42	.36
Full	6.4	3.5	4.3	2.7	11.0	4.8
	1.8	1.1	1.4	.94	2.9	1.3
	.69	.52	.55	.45	.90	.50

Jeffreys' priors I

- for $Y_i \sim N(0, \mathbf{R})$, Jeffreys' prior is

$$p(\mathbf{R}) \propto |\mathbf{R}|^{-(p+1)/2}$$

- we can rewrite this as a function of the partial autocorrelations using the following determinant identity (Joe, 2006; Daniels and Pourahmadi, 2009)

$$|\mathbf{R}| = \prod_{j=2}^p \prod_{k=1}^{j-1} (1 - \pi_{j,k}^2)$$

Jeffreys' priors II

- recall, the jacobian from \mathbf{R} to $\boldsymbol{\pi}$ is

$$J(\mathbf{R} \rightarrow \boldsymbol{\pi}) = \prod_{k=1}^{p-1} \prod_{i=1}^{p-k} (1 - \pi_{i,i+k}^2)^{(p-1-k)/2}$$

- so

$$\rho(\boldsymbol{\pi}) \propto \prod_{k=1}^{p-1} \prod_{i=1}^{p-k} \left(\frac{1}{(1 - \pi_{i,i+k}^2)} \right)^{(k+2)/2}$$

- this is the opposite behavior that we want (as with Barnard Beta prior)
 - as lag increases, more mass away from zero
 - not surprising as these priors are 'ignorant' of the ordering and behave this way on partial correlations to obtain identical marginal priors for the marginal correlations

Marginally uniform prior I

$$p(\mathbf{R}) \propto |\mathbf{R}|^{(p-1)(p+1)/2-2} \left(\prod_i r^{ii} \right)^{-(p+1)/2}$$

where r^{ii} is the i th diagonal element of \mathbf{R}^{-1} (i.e., the partial variance)

- derivation based on marginalized of Wishart prior with scale matrix \mathbf{I} .

$$p(\boldsymbol{\pi}) \propto \frac{\prod_i \prod_k (1 - \pi^2_{i, i+k})^{(p-1)(p+1)/2-2+(p-1-k)/2}}{\left(\prod_i r^{ii} \right)^{-(p+1)/2}}$$

Marginally uniform prior II

- similar behavior on partial correlations; more weight away from zero as lag increases (opposite of preferred behavior)
- can't express as independent (Beta) priors on partial correlations (for $p \geq 4$) (as with Jeffreys')
- not easily adaptable to structural zeros (as with Jeffreys')

Class of Priors I

- Use independent (linearly transformed) Beta priors on the interval $(-1,1)$ for partial correlations
- convenient and flexible way to specify a prior for a correlation matrix R
- Barnard Beta and independent uniform are special cases
- easy to adapt to structural zeros
- invariant to sequentially adding to the dimension
- shrinkage priors similar to Daniels and Kass (2001) and Daniels and Pourhamdi (2002) based on these Beta priors, e.g.,

$$\pi_{j,j+k} \sim \text{Beta}_{[-1,1]}(\pi_k^*, \delta) : E[\pi_{j,j+k}] = \pi_k^*$$

Posterior computations: Unstructured \mathbf{R}

- use parameter expanded Metropolis-Hastings
- details in Liu and Daniels (2006; JCGS); extends previous work by Liu (2001) to any prior
- idea is to expand to a covariance matrix, sample from an inverse Wishart using a determined prior, and then transform back and accept based on ratio of determined prior to prior of interest (Metropolis-Hastings step)

Posterior computations: Structural zeros

- with zero partial correlations, auxiliary variable sampler (Damien, Wakefield, and Walker 1999) works quite well

$$\int I\{U_{jk} < \text{Lik}(\boldsymbol{\pi})\} p(\boldsymbol{\pi}_{jk}) dU_{jk}$$

- easy to implement; each partial correlation free to vary independently in bounded interval, $[-1, 1]$ - works well, low autocor in chain
- within class of independent Beta priors, just need to sample from a uniform and a truncated Beta distribution (easy; Damien and Walker, 2002)

Posterior Computations: Useful facts

- Fact 1.** If we factor $R^{-1} = CPC$, where P is a correlation matrix and C is a diagonal matrix, the elements of P are the full partial correlations, $\text{corr}(Y_j, Y_k | \{Y_l : l \neq j, l \neq k\})$
- Fact 2.** To isolate the likelihood contribution of $\pi_{j,j+k}$, we can factor the entire multivariate normal distribution into

$$p(y_j, \dots, y_{j+k})p(y_l : l < j \text{ or } l > j + k \mid y_j, \dots, y_{j+k})$$

The (l, k) entry of inverse of the correlation matrix, $R[j : j + k]$ for the first factor is related to the partial autocorrelation of interest

Ongoing work

- (reference) priors; add in partial info: $\pi_1 \geq \pi_2 \geq \dots \geq \pi_{p-1}$
- dimension reduction
 - set all partial autocorrelations greater than lag k ($k \ll p$) to 0; model the rest
 - reduces to only dealing with correlation matrices of dimension k
 - can sequentially compute the MLE in closed form when any subset of the π 's are zero (just solving cubics)
- efficient computations with covariates
- automated parsimonious modelling (more next)
- ordering - Graphical models - for decomposable graphs, connections between zeros in covariance matrix and Choleski factor, ?

Automated Parsimonious modeling

- adapt automated 'zeroing' out from Wong, Carter, and Kohn (2003) [WCK]
 - prior with mixture of non-degenerate distribution and point mass at zero
 - great simplifications in our setting with partial autocorrelations (they used full partial correlations; constraints are quite problematic)
 - details next

Parsimonious modeling

- Let A_{jk} be an indicator that π_{jk} is non-zero (and \mathbf{A} the vector of indicators for all the π_{jk})
- Let S be the number of non-zero partial correlations, $\{0, \dots, p(p-1)/2\}$
- specify

$$p(d\pi|\mathbf{A})p(\mathbf{A}|S)p(S)$$

Specification of $p(d\boldsymbol{\pi} | \mathbf{A})$

- Similar to WCK, we can specify a uniform prior on the partial correlations for a specific pattern, \mathbf{A} ,

$$p(d\boldsymbol{\pi} | \mathbf{A}) = \frac{1}{V(\mathbf{A})} d\boldsymbol{\pi}_{\mathbf{A}=1}$$

where $V(\mathbf{A})$ is the normalizing constant and $\mathbf{A} = 1$

- corresponds to the prior for the non-zero π_{jk} ; just the uniform prior on $\boldsymbol{\pi}$ from earlier.

Computation of $V(\mathbf{A})$

- here,

$$V(\mathbf{A}) = 2^{S(\mathbf{A})}$$

where $S(\mathbf{A})$ is the number of non-zero partial correlations

- simple since the partial correlations are free to vary in $[-1, 1]$ for any positive definite correlation matrix \mathbf{R} (just a $S(\mathbf{A})$ -dimensional hypercube as earlier)
- corresponding prior in WCK for the full partial correlations is a highly constrained space since the full partial correlations are *not unconstrained*
 - normalizing constants, $V(\mathbf{A})$ need to be computed using Monte Carlo methods (and priors even specified in certain ways to minimize this computational burden)
- easily adapt to other priors, $p(d\pi|\mathbf{A})$, e.g., independent Beta priors

Specification of $p(\mathbf{A}|S)$

- The second component of the prior is the distribution of the configuration of non-zero partial correlations given that there are $S(\mathbf{A}) = l$ non-zero partial correlations, $p(\mathbf{A} | S(\mathbf{A}) = l)$.
- Given a particular number of non-zero correlations, WCK consider any set of l partial correlations being non-zero equally likely.
- In our setting, we do not expect such *exchangeability*.
- will use a prior that allows lag 1 partial correlations to be more likely to be non-zero and higher order lag partial correlations less likely to be non-zero.
 - a separate probability, $\psi_k = P(\text{lag } k \text{ non-zero})$ for each set of lag k partial correlation such that $\psi_1 > \psi_2 > \dots > \psi_{p-1}$.

Specification of $p(S)$

- The final components of the prior is the distribution of the number of non-zero partial correlations, $S(\mathbf{A})$.
- This is the piece of the prior that determines the overall parsimony.
- specify this prior based on prior belief about the expected sparsity of π . (For example, an AR(1) type structure would correspond to only the lag 1 partial correlations being nonzero)

Pending: Computations

- how best to do the sampling?
 - given structure of partial correlations (and bounded domain), possibly sequential importance sampling ideas (Moral, Doucet, Jasra, 2006)
 - preliminary calculations suggest partial autocorrelations have lower pairwise correlations than marginal correlations; useful to exploit for one at a time (Gibbs) sampling
- extension to allow point mass for stationarity within lag, $\pi_{j,j+k} = \pi_k^*$ (more complex prior)
- Dirichlet priors on π 's to do flexible grouping