Generating random correlation matrices based on partial correlation vines and the onion method

Harry Joe
Department of Statistics
University of British Columbia

Abstract

Partial correlation vines and the onion method are presented for generating random correlation matrices. As a special case, a uniform distribution over the set of $d \times d$ positive definite correlation matrices obtains.

Byproducts are: (a) For a uniform distribution over the space of $d \times d$ correlation matrices, the marginal distribution of each correlation is Beta$(d/2, d/2)$ on $(-1, 1)$. (b) An identity is obtained for the determinant of a correlation matrix R via partial correlations in a vine. (c) A formula is obtained for the volume of the set of $d \times d$ positive definite correlation matrices in $\binom{d}{2}$-dimensional space.

Outline

1. Statement of key results on generating random correlation matrices
2. Parametrization with partial correlations
3. Regular vines, D-vines, C-vines
4. Random correlation matrices based on partial correlations
5. Random correlation matrices based on onion method

Key results: R is a $d \times d$ correlation matrix

1. Several simple ways to generate a random R that is uniform over set of positive definite $d \times d$ correlation matrices; more generally with density $\propto [\det(R)]^{\alpha-1}$.

2. Identity: $(1-\rho_{12}^2)(1-\rho_{23}^2)(1-\rho_{13;2}^2)$ for $d = 3$

$$\det(R) = \prod_{i=1}^{d-1}(1-\rho_{i,i+1}^2) \times \prod_{k=2}^{d-1}\prod_{j=1}^{d-k}(1-\rho_{j,j+k;j+1\ldots j+k-1}^2).$$

[partial correlations in the double product]

3. Volume of $d \times d$ correlation matrices in $2^{d(d-1)/2}$ dimensional space is:

$$\begin{cases} \pi^{(d^2-1)/4}\ \dfrac{\prod_{m=1}^{(d-1)/2}\Gamma(2m)}{2^{(d-1)^2/4}\,\Gamma^{d-1}\left(\frac{d+1}{2}\right)}, & \text{if } d \text{ is odd;} \\ \pi^{d(d-2)/4}\ \dfrac{2^{(3d^2-4d)/4}\,\Gamma^{d}\left(\frac{d}{2}\right)\prod_{m=1}^{(d-2)/2}\Gamma(2m)}{\Gamma^{d-1}(d)}, & \text{if } d \text{ is even.} \end{cases}$$

Partial correlations

correlations $\rho_{i,i+1}$ for $i = 1, \ldots, d-1$
the partial correlations $\rho_{ij;i+1,\ldots j-1}$ for $j - i \geq 2$
For example, $\rho_{12}, \rho_{23}, \rho_{34}, \rho_{13;2}, \rho_{24;3}, \rho_{14;23}$ for $d = 4$.

For multivariate normal random variables, partial correlations are conditional correlations. In general, they are functions of a correlation matrix.

Advantage: the $\binom{d}{2}$ parameters can independently take values in the interval $(-1, 1)$; different vines can be used.

Disadvantage: the reparametrization is non-unique, and depends on the order of indexing, e.g.: $\rho_{12}, \rho_{13}, \rho_{23;1}$ or $\rho_{12}, \rho_{23}, \rho_{13;2}$

$d = 5$: 3 different vines + matrix for onion method are shown.

d-vine which is like Markovian dependence | c-vine

regular partial correlation vine | onion method

#distinct vines increases quickly as d increases,
partial correlations independently in $(-1, 1)$.

$$\begin{pmatrix} 1 & \rho_{12} & \rho_{13;2} & \rho_{14;23} & \rho_{15;234} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24;3} & \rho_{25;34} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} & \rho_{35;4} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} \\ \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} \\ \rho_{12} & 1 & \rho_{23;1} & \rho_{24;1} & \rho_{25;1} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34;12} & \rho_{35;12} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45;123} \\ \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & \rho_{12} & \rho_{13;2} & \rho_{14;2} & \rho_{15;24} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} & \rho_{25;4} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34;12} & \rho_{35;124} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} \\ \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} & \rho_{25} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} & \rho_{35} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} \\ \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & 1 \end{pmatrix}$$

D-vine for $d = 5$: specify densities separately for each partial correlation; then density of the correlation matrix $\mathbf{R}$ is

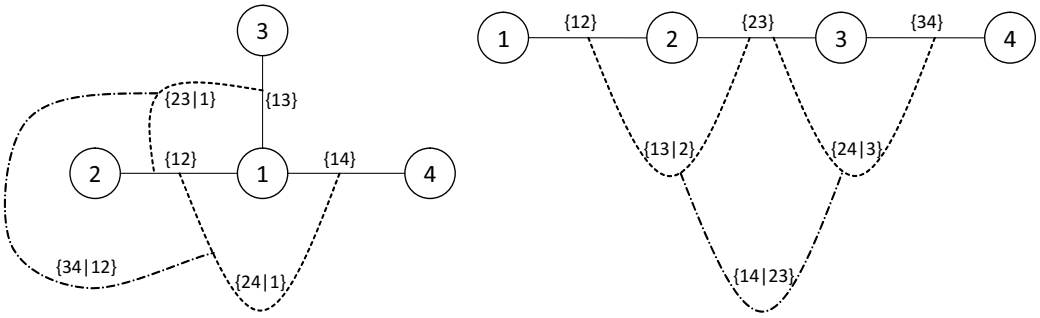
$$\begin{aligned} f_{\mathbf{R}}(\mathbf{r}) = \quad & f_{\rho_{12}}(r_{12}) f_{\rho_{23}}(r_{23}) f_{\rho_{34}}(r_{34}) f_{\rho_{45}}(r_{45}) \\ \times \quad & f_{\rho_{13;2}}(r_{13;2}) f_{\rho_{24;3}}(r_{24;3}) f_{\rho_{35;4}}(r_{35;4}) \\ \times \quad & f_{\rho_{14;23}}(r_{14;23}) f_{\rho_{25;34}}(r_{25;34}) f_{\rho_{15;234}}(r_{15;234}) \\ \times \quad & \text{Jacobian} \end{aligned}$$

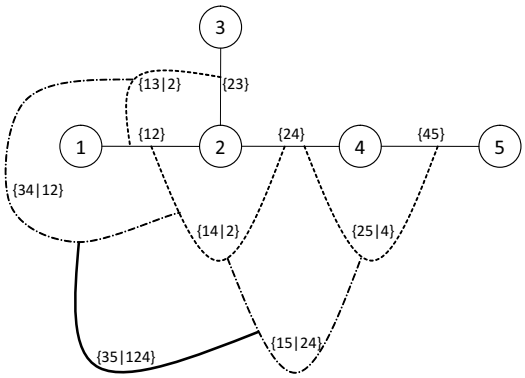
Under appropriate choices, this leads to a uniform density.

References for vines

D. Kurowicka, R. Cooke, *Uncertainty Analysis with High Dimensional Dependence Modelling*, Wiley, 2006.

Vines are a graphical method for modeling multivariate dependencies





Random correlation matrices [Joe, 2006, JMVA]

Partial correlation approach: distributions separately for $\rho_{i,i+1}$ for $i = 1, \ldots, d-1$, and $\rho_{ij;i+1,\ldots j-1}$ for $j - i \geq 2$.

This leads to a joint distribution of $\{\rho_{ij} : 1 \leq i < j \leq d\}$.

More generally, choose distributions separately for partial correlation in a regular vine.

Previous papers : combination of random eigenvalues and random (orthogonal) matrices, or a random T matrix leading to $T'T$, but cannot do distribution theory?

Motivations:

- generation of ellipsoidal clusters to study clustering methods
- efficiency of estimating eq'n methods for multiv probit model
- assess probability that a correlation matrix is a rank correlation matrix.
- prior for correlation matrix in Bayesian statistics?

Main result

With appropriate choices of beta distributions for $\rho_{i,i+1}$ for $i = 1, \ldots, d-1$, and $\rho_{ij;i+1,\ldots j-1}$ for $j - i \geq 2$, can get a joint density for $\{\rho_{ij} : 1 \leq i < j \leq d\}$ that is $\propto |R|^{\alpha-1}$, where $\alpha > 0$.

More generally, same for partial correlations on a regular vine.

In this case, the joint density is **invariant** to the order of indexing of variables for the partial correlations (and to the vine), and each ρ_{ij} marginally has a Beta$(\alpha - 1 + d/2, \alpha - 1 + d/2)$ distribution on $(-1, 1)$.

Special case with $\alpha = 1$: uniform density over the set of positive definite correlation matrices.

Anderson (1958, p. 80): partial corr $\rho_{j,j+k;j+1,\ldots,j+k-1}$ $(2 \leq k \leq d-1)$ is

$$\frac{\rho_{j,j+k} - \mathbf{r}_1'(j,k)(R_2(j,k))^{-1}\mathbf{r}_3(j,k)}{\left[1 - \mathbf{r}_1'(j,k)(R_2(j,k))^{-1}\mathbf{r}_1(j,k)\right]^{1/2}\left[1 - \mathbf{r}_3'(j,k)(R_2(j,k))^{-1}\mathbf{r}_3(j,k)\right]^{1/2}}$$

where

$$R[j : j+k] = \begin{pmatrix} 1 & \mathbf{r}_1'(j,k) & \rho_{j,j+k} \\ \mathbf{r}_1(j,k) & R_2(j,k) & \mathbf{r}_3(j,k) \\ \rho_{j+k,j} & \mathbf{r}_3'(j,k) & 1 \end{pmatrix},$$

with $\mathbf{r}_1'(j,k) = (\rho_{j,j+1}, \ldots, \rho_{j,j+k-1})$, $\mathbf{r}_3'(j,k) = (\rho_{j+k,j+1}, \ldots, \rho_{j+k,j+k-1})$, and $R_2(j,k)$ consisting of the middle $k-1$ rows / columns. Hence

$$\rho_{j,j+k} = \mathbf{r}_1'(j,k)(R_2(j,k))^{-1}\mathbf{r}_3(j,k) + \rho_{j,j+k;j+1,\ldots,j+k-1}D_{jk},$$

where

$$D_{jk}^2 = \Big[1 - \mathbf{r}_1'(j,k)(R_2(j,k))^{-1}\mathbf{r}_1(j,k)\Big]\Big[1 - \mathbf{r}_3'(j,k)(R_2(j,k))^{-1}\mathbf{r}_3(j,k)\Big].$$

$d = 3$ [ρ for rv, r for dummy variable of function]
Consider $(\rho_{12}, \rho_{23}, \rho_{13\cdot 2}) \to (\rho_{12}, \rho_{23}, \rho_{13})$. Since

$$\rho_{13\cdot 2} = \frac{\rho_{13} - \rho_{12}\rho_{23}}{\sqrt{(1-\rho_{12}^2)(1-\rho_{23}^2)}},$$

the Jacobian of the transformation is:

$$\partial r_{13\cdot 2}/\partial r_{13} = [(1-r_{12}^2)(1-r_{23}^2)]^{-1/2}.$$

Note that

$$1 - r_{13\cdot 2}^2 = \frac{1 - r_{12}^2 - r_{23}^2 - r_{13}^2 + 2r_{12}r_{23}r_{13}}{(1-r_{12}^2)(1-r_{23}^2)} = \frac{\det(R)}{(1-r_{12}^2)(1-r_{23}^2)}$$

$$\det(R(r_{12}, r_{23}, r_{23})) = 1 - r_{12}^2 - r_{23}^2 - r_{13}^2 + 2r_{12}r_{23}r_{13}$$

$$= (1-r_{12}^2)(1-r_{23}^2)(1-r_{13\cdot 2}^2) \qquad (D1)$$

Beta(α,α) density on $(-1,1)$: $\frac{1}{2}[B(\alpha,\alpha)]^{-1}(1-u^2)^{\alpha-1}$
or $U = 2V - 1$, $V \sim$ Beta(α,α) on $(0,1)$.

$\rho_{12}, \rho_{23} \sim$ Beta(α_1, α_1),
$\rho_{13\cdot 2} \sim$ Beta(α_2, α_2) independently.

$$f_{\rho_{12},\rho_{23},\rho_{13}}(r_{12}, r_{23}, r_{13}) \propto (1-r_{13\cdot 2}^2)^{\alpha_2 - 1}[(1-r_{12}^2)(1-r_{23}^2)]^{\alpha_1 - 3/2}$$

If $\alpha_1 = \alpha_2 + \frac{1}{2}$, then the density is

$$\propto [\det(R)]^{\alpha_2 - 1} \qquad (D2)$$

This is symmetric in r_{12}, r_{13}, r_{23}, which means the same joint density obtains from $\rho_{12}, \rho_{13} \sim$ Beta(α_1,α_1), $\rho_{23\cdot 1} \sim$ Beta(α_2,α_2) etc.

$\alpha_2 = 1$, $\alpha_1 = 3/2 \Rightarrow$ Uniform over set of correlation matrices.

$d > 3$: computer proof before math proof

Based on the results for the $d = 3$ case, conjectured an extension of (D1) as an identity for $\det(R)$, and conjectured the Beta distributions on $\rho_{ij;i+1,\ldots,j-1}$ needed to get joint density of (ρ_{ij}) to be $\propto [\det(R)]^{\alpha-1}$.

The identity was verified numerically and the univariate margins of the random correlation matrix were also checked numerically before proving the results for general d.

Results for $d > 3$.
Thm 1:

$$\det(R) = \prod_{i=1}^{d-1}(1-\rho_{i,i+1}^2) \times \prod_{k=2}^{d-1}\prod_{j=1}^{d-k}(1-\rho_{j,j+k;j+1\ldots j+k-1}^2).$$

Thm 4: The determinant $|J_d|$ of the Jacobian for the transform of $(\rho_{12}, \rho_{23}, \rho_{13}, \rho_{34}, \rho_{24}, \rho_{14}, \rho_{45}, \ldots, \rho_{1d})$ to $(\rho_{12}, \rho_{23}, \rho_{13;2}, \rho_{34}, \rho_{24;3}, \rho_{14;23}, \rho_{45}, \ldots, \rho_{1d;2\ldots d-1})$ is:

$$\Big[\prod_{i=1}^{d-1}(1-\rho_{i,i+1}^2)^{d-2} \times \prod_{k=2}^{d-2}\prod_{i=1}^{d-k}(1-\rho_{i,i+k;i+1\ldots i+k-1}^2)^{d-1-k}\Big]^{-1/2}.$$

Thm 5: If $\alpha_k = \alpha_{d-1} + \frac{1}{2}(d-1-k)$, $k = 1, \ldots, d-1$, and $\rho_{i,i+k;i+1\ldots i+k-1}$ is Beta(α_k, α_k) on $(-1,1)$ for $1 \le i < i+k \le d$, then the joint density becomes

$$c_d^{-1}\left[\det\{(r_{ij})_{1\le i,j\le d}\}\right]^{\alpha_{d-1}-1},$$

where the normalizing constant c_d is

$$2^{\sum_{k=1}^{d-1}(2\alpha_{d-1}-2+d-k)(d-k)} \times \prod_{k=1}^{d-1}\left[B\left(\alpha_{d-1}+\tfrac{1}{2}(d-1-k),\, \alpha_{d-1}+\tfrac{1}{2}(d-1-k)\right)\right]^{d-k}.$$

If $\alpha_{d-1} = 1$ and $\alpha_k = \frac{1}{2}(d+1-k)$ for $k = 1, \ldots, d-2$, leading to uniform joint density for $\{\rho_{ij}, i<j\}$, then the normalizing constant is

$$c_d = 2^{\sum_{k=1}^{d-1}(d-k)^2} \times \prod_{k=1}^{d-1}\left[B\left(\tfrac{1}{2}(d-k+1), \tfrac{1}{2}(d-k+1)\right)\right]^{d-k}$$

$$= 2^{\sum_{k=1}^{d-1}k^2} \times \prod_{k=1}^{d-1}\left[B\left(\tfrac{1}{2}(k+1), \tfrac{1}{2}(k+1)\right)\right]^{k},$$

and the recursion is

$$c_d = c_{d-1} \times 2^{(d-1)^2} \times \left[B(\tfrac{1}{2}d, \tfrac{1}{2}d)\right]^{d-1}.$$

Byproduct: normalizing constant is the volume of the set of d-dim positive definite correlation matrices in $\binom{d}{2}$-dim space

d	c_d	$c^d/2^{d(d-1)/2}$
2	2	
3	$4.934802 = \pi^2 \cdot (1/2)$	0.617
4	$11.69731 = \pi^2 \cdot (32/27)$	0.183
5	$22.53256 = \pi^6 \cdot (3/128)$	0.022
6	$31.11388 = \pi^6 \cdot (8192/253125)$	0.00095
7	$27.85823 = \pi^{12} \cdot (n_7/d_7)$	0.000013
8	$14.87740 = \pi^{12} \cdot (n_8/d_8)$	5.5×10^{-8}
9	$4.411544 = \pi^{20} \cdot (n_9/d_9)$	6.4×10^{-11}
10	$0.682269 = \pi^{20} \cdot (n_{10}/d_{10})$	1.9×10^{-14}

$c^d/2^{d(d-1)/2}$ decreases quickly as d increases.

Random correlation matrix based on vines [Lewandowski, Kurowicka and Joe, 2007]

Extensions of Theorems 1,4,5 in Joe (2006, JMVA).

$V = (e)$ is a partial correlation vine;
for example, $V = (12, 13, 14, 23; 1, 24; 1, 34; 12)$ with
depth=(order-1)=(#conditioned variables): $(0, 0, 0, 1, 1, 2)$.
For a vine based on d variables, there are $d-1$ edges with depth $k_e = 0$, $d-2$ edges with depth $k_e = 1$, ..., and 1 edge with depth $k_e = d-2$.

Let R be a (random) correlation matrix based on distributions for $\{\rho_e : e \in V\}$.

T1: $\det(R) = \prod_{e\in V}(1-\rho_e^2)$ [Kurowicka and Cooke, 2006, LAA] (product over edges of the vine)

T4: Jacobian of transform from correlations $\{\rho_{ij}\}$ to partial correlations $\{\rho_e\}$ is

$$\left\{\prod_{e\in V}(1-\rho_e^2)^{d-2-k_e}\right\}^{-1/2}$$

T5. Let $\rho_e \sim \text{Beta}(\alpha_{k_e+1}, \alpha_{k_e+1})$ on $(-1,1)$ independently, where $\alpha_{k_e+1} = \alpha_{d-1} + \frac{1}{2}(d-2-k_e)$. Get uniform over space of correlation matrices if $\alpha_{d-1} = 1$, and then marginal distribution for each correlation is Beta($d/2$, $d/2$).

The C-vine algorithm for generating correlation matrices with density $\propto [\det(\mathbf{r})]^{\eta-1}$, $\eta > 1$.

1. Initialization $\beta = \eta + (d-1)/2$.
2. Loop for $k = 1, \ldots, d-1$.
   a) $\beta \leftarrow \beta - \frac{1}{2}$;
   b) Loop for $i = k+1, \ldots, d$;
      i) generate partial corr $p_{k,i;1,\ldots,k-1} \sim \text{Beta}(\beta,\beta)$ on $(-1,1)$;
      ii) use recursive formula on $p_{k,i;1,\ldots,k-1}$ to get $r_{k,i} = r_{i,k}$ [no matrix inversions];
3. Return $\mathbf{r}$, a $d \times d$ correlation matrix.

Onion method for random correlation matrices
[Ghosh & Henderson, 2003; extended LKJ 2007]

Result 1. Consider the spherical density $c(1-\mathbf{w}'\mathbf{w})^{\beta-1}$ for $\mathbf{w} \in \Re^m$, $\mathbf{w}'\mathbf{w} \le 1$, where $c = \Gamma(\beta+m/2)\pi^{-m/2}/\Gamma(m/2)$. If $\mathbf{W}$ has this density, then $\mathbf{W} = V\mathbf{U}$ where $V \sim \text{Beta}(m/2, \beta)$ and $\mathbf{U}$ is uniform on the surface of the m-dimensional hypersphere. If $\mathbf{Q} = \mathbf{AW}$, where $\mathbf{A}$ is an $m \times m$ non-singular matrix, then the density of $\mathbf{Q}$ is

$$c[\det(\mathbf{AA}')]^{-1/2}(1-\mathbf{q}'[\mathbf{AA}']^{-1}\mathbf{q})^{\beta-1}, \quad \mathbf{q} \ni \mathbf{q}'[\mathbf{A}'\mathbf{A}]^{-1}\mathbf{q} \le 1.$$

Result 2. Partition $\mathbf{r}_{m+1} = \begin{pmatrix} \mathbf{r}_m & \mathbf{q} \\ \mathbf{q}' & 1 \end{pmatrix}$ where $\mathbf{r}_m$ is an $m \times m$ correlation matrix and $\mathbf{q}$ is a m-vector of correlations such that $\mathbf{r}_{m+1}$ is an $(m+1) \times (m+1)$ correlation matrix. Then

$$\det(\mathbf{r}_{m+1}) = \det(\mathbf{r}_m) \cdot \left(1 - \mathbf{q}'\mathbf{r}_m^{-1}\mathbf{q}\right).$$

We use upper case letter of $\mathbf{r}_m, \mathbf{q}, \mathbf{r}_{m+1}$ to denote random vectors and matrices. Let $\beta, \beta_m > 0$. If $\mathbf{R}_m$ has density $\propto [\det(\mathbf{r}_m)]^{\beta_m - 1}$ and
$\mathbf{Q}$ given $\mathbf{R}_m = \mathbf{r}_m$ has density $\propto [\det(\mathbf{r}_m)]^{-1/2}(1-\mathbf{q}'\mathbf{r}_m^{-1}\mathbf{q})^{\beta-1}$,
then the density of $\mathbf{R}_{m+1}$ is $\propto [\det(\mathbf{r}_m)]^{\beta_m - 3/2}(1-\mathbf{q}'\mathbf{r}_m^{-1}\mathbf{q})^{\beta-1}$.
If $\beta_m = \beta + \frac{1}{2}$, then the density of $\mathbf{R}_{m+1}$ is $\propto [\det(\mathbf{r}_{m+1})]^{\beta-1}$.

Algorithm for the extended onion method to get random correlation matrices in dimension d with density $\propto [\det(\mathbf{r})]^{\eta-1}$, $\eta > 0$.

1. Initialization. $\beta = \eta + (d-2)/2$, $r_{12} \leftarrow 2u - 1$, where $u \sim \text{Beta}(\beta,\beta)$, $\mathbf{r} \leftarrow \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$
2. Loop for $m = 2, \ldots, d-1$.

(a) $\beta \leftarrow \beta - \frac{1}{2}$;

(b) generate $y \sim \text{Beta}(m/2, \beta)$, generate $\mathbf{u} = (u_1, \ldots, u_m)'$ uniform on the surface of m-dimensional hypersphere;

(c) $\mathbf{w} \leftarrow y^{1/2}\mathbf{u}$, obtain Cholesky decomp $\mathbf{AA}' = \mathbf{r}$, set $\mathbf{q} \leftarrow \mathbf{Aw}$;

(d) $\mathbf{r} \leftarrow \begin{pmatrix} \mathbf{r} & \mathbf{q} \\ \mathbf{q}' & 1 \end{pmatrix}$.

3. Return $\mathbf{r}$, a $d \times d$ correlation matrix.

Why is this called the *onion* method ???

By the symmetry, each $\rho_{jk} = R_{jk}$ in the correlation matrix has a marginal Beta($\eta + [d-2]/2$, $\eta + [d-2]/2$) density on $(-1, 1)$. For the special case of $\eta = 1$ leading to uniform over the space of correlation matrices, the marginal density of each R_{jk} is Beta($d/2, d/2$) on $(-1, 1)$.

Computational time:

In C, onion method with incremental Cholesky is fastest, and C-vine is faster than onion method without incremental Cholesky.
In Matlab, C-vine (which avoids matrix inversions) is fastest.
D-vine is slower because of matrix inversions.

Other comments:

If partial correlations in a particular vine are needed, then use the appropriate vine.
Choosing the vine and the densities for the partial corr in this vine, one could get random correlation matrices that have larger correlations at a few particular pairs.

References

H. Joe, Generating random correlation matrices based on partial correlations, Journal of Multivariate Analysis 97 (2006) 2177–2189.

S. Ghosh, S.G. Henderson, Behavior of the NORTA method for correlated random vector generation as the dimension increases. ACM Transactions on Modeling and Computer Simulation (TOMACS) 13 (2003), 276–294.

T. Bedford, R. Cooke, Vines - a new graphical model for dependent random variables, Annals of Statistics 30 (2002) 1031–1068.

D. Kurowicka, R. Cooke, *Uncertainty Analysis with High Dimensional Dependence Modelling*, Wiley, 2006.

D. Kurowicka, R. M. Cooke, Completion problem with partial correlation vines, Linear Algebra and Its Applications 418 (2006) 188–200.

D. Lewandowski, D. Kurowicka, H. Joe, Generating random correlation matrices based on vines and extended Onion method, (2007), submitted.