JDSSV

Journal of Data Science, Statistics, and Visualisation

MMMMMM YYYY, Volume VV, Issue II. *doi: XX.XXXXX/jdssv.v000.i00*

Visualizing Distributions of Covariance Matrices

Tomoki Tokuda
University of Leuven, Belgium
ERI/University of Tokyo, Japan

Ben Goodrich
Columbia University, New York, USA

Iven Van Mechelen
University of Leuven, Belgium

Andrew Gelman
Columbia University, New York, USA

Francis Tuerlinckx
University of Leuven, Belgium

Abstract

Statistical graphics are generally designed for visualizing data, but in this case our primary goal is to understand complex multivariate distributions that might be used as prior distributions for models with unknown covariance matrices. Visualizing a distribution of covariance matrices is a step beyond visualizing a single covariance matrix or a single multivariate dataset. We take advantage of the symmetries in many standard prior distributions to efficiently and effectively display these highly multivariate distributions using a tableau of low-dimensional displays. We demonstrate our approach for graphing distributions of covariance matrices on several models, including the Wishart, inverse-Wishart, and scaled inverse-Wishart families in different dimensions. Our visualizations follow the principle of decomposing a covariance matrix into scale parameters and correlations, pulling out marginal summaries where possible and using two- and three-dimensional plots to reveal multivariate structure. Our visualization methods are available through the R package **VisCov**.

Keywords: Bayesian statistics, prior distributions, Wishart distribution, inverse-Wishart distribution, LKJ distribution, statistical graphics.

1. Background

Covariance matrices and their corresponding distributions play an important role in statistics. Graphical visualizations can be used to understand probability distributions. But visualizing a distribution in a high-dimensional space is a challenge: Traditional tools such as plots of densities and distributions, contour plots, and point clouds become less useful in high dimensions. With covariance matrices, there is the additional difficulty that they must be positive semi-definite, a restriction that forces the joint distribution of the covariances into an oddly-shaped subregion of the space.

Distributions of covariance matrices show up in frequentist (e.g., Anderson 2003) and Bayesian statistics (e.g., Yang and Berger 1994; Daniels and Kass 1999; Barnard et al. 2000):

- The sampling distribution of the covariance matrix of independent multivariate observations: If the data are generated according to a multivariate normal distribution, then their covariance matrix has a Wishart sampling distribution (see Wishart 1928).
- The prior for a covariance matrix in a Bayesian analysis, most simply if data are modeled as independent draws from a multivariate normal with an unknown mean and covariance matrix; in the same vein, a prior on the residual covariance matrix is needed in a multivariate linear regression model (Box and Tiao 1973; Zellner 1971).
- In hierarchical regression model with varying intercepts and slopes, the vector of varying coefficients for each group is often assumed to follow a multivariate normal distributed random variable with mean zero and an unknown covariance matrix, which again requires a prior distribution in a Bayesian analysis (see, e.g., Gelman and Hill 2007).
- The covariance matrix itself may be unit-specific and drawn independently from a population distribution of covariance matrices; such a situation is less common but can occur in both frequentist and Bayesian settings (see Oravecz et al. 2009; De Boeck et al. 2024).

In the remainder of this paper, we aim to visualize distributions of covariance matrices, while focusing on such distributions when used as priors in a Bayesian analysis. Indeed, such an analysis often requires choosing a family of prior distributions or understanding a prior distribution that had already been specified before. Several classes of priors for covariance matrices have been proposed in the statistical literature, but many of their properties are unknown analytically and, in general, extensive expertise with these priors has not yet been acquired (see also Merkle et al. 2023).

As an example, consider the inverse-Wishart distribution, which is often used in Bayesian analyses because it is a proper conjugate prior for an unknown covariance matrix in a multivariate normal model (see Gelman et al. 2013). Some specific analytical results for the inverse-Wishart have been derived. For example, the marginal distribution of a diagonal block submatrix of draws from an inverse-Wishart distribution is also

inverse-Wishart (Press 1982). Marginal moments of such draws have been derived as well (Von Rosen 1988). But marginal distributions are typically not known and there is no expression for the bivariate distribution of any two covariances. As a result, analytical knowledge of the properties of the inverse-Wishart distribution is still highly incomplete. Various alternatives to the inverse-Wishart have been proposed (Barnard et al. 2000; O'Malley and Zaslavsky 2008; Lewandowski et al. 2009), but still fewer analytical results are known for them, making it even more challenging to understand precisely their properties. In sum, our analytical understanding of these distributions falls short of providing us a full understanding of them.

Since analytical results are limited, other tools are needed to study covariance matrix distributions. To obtain insight into the properties of such distributions, visualization might be considered. There is a considerable literature concerning visualization of multivariate data (e.g., Valero-Mora et al. 2003; Theus and Urbanek 2008; Cook and Swayne 2007). For visualization of a covariance or correlation matrix, in particular, a heatmap is often used (e.g., Friendly 2002). Yet, this and other existing approaches tend to focus on the visualization of a *single* covariance matrix (derived from a single data set), while we would like to visualize *distributions* of covariance matrices. For the latter, there is a need for specialized methods, because not all techniques carry over easily from the single instance to the distribution case. For instance, averaging a heatmap over a number of instances of correlation matrices ends up with displaying the mean correlation matrix, which does not capture the variability of the distribution.

As a way out, we propose a series of graphs to visualize covariance matrix distributions. To cope with the high dimensionality of the distributions in question, we will visualize key aspects of them while relying as much as possible on exchangeability (i.e., invariance under permutations of variables) in conventional classes of prior distributions of covariance matrices. For any joint distribution of the covariances, we construct a grid showing the marginal distribution of the scale and correlation parameters, along with two- and three-dimensional scatterplots (e.g., Liu et al. 2016; Peters et al. 2014; Alvarez et al. 2014). Since a covariance matrix can be expressed as an equiprobability or isodensity ellipse in case of a multivariate normal distribution (or, more generally, in case of any elliptical distribution, including, e.g., multivariate t-distributions; Owen and Rabinovitch 1983), we also display the distribution as a mixture of ellipses in a multivariate normal setting. We demonstrate for several distributions—including the Wishart, inverse-Wishart, scaled inverse-Wishart, and uniform correlation—that this series of graphs allows salient features to be visualized, analyzed, and compared. Our ultimate goal is to offer a method and tool to help us better understand the multivariate distributions under study.

2. A four-layer visualization method

Let us first introduce some notation. A $k \times k$ covariance matrix $\mathbf{\Sigma}$ has variances σ_i^2 $(i = 1, \ldots, k)$ on its diagonal. The typical off-diagonal element is $\sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$ $(i = 1, \ldots, k, j = 1, \ldots, k, i \neq j)$, where σ_i is the standard deviation of the ith variable and ρ_{ij} the correlation between the ith and jth variables. We often separate covariances into standard deviations and scale-free correlations for interpretability.

We begin with the inverse-Wishart distribution:

$$\boldsymbol{\Sigma} \sim \text{inv-Wishart}_{\nu}(\boldsymbol{S}^{-1}), \tag{1}$$

where ν denotes the degrees of freedom and $\boldsymbol{S}$ is a positive definite $k \times k$ scale matrix. The density of $\boldsymbol{\Sigma}$ is proportional to:

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+k+1)/2} \exp\left(-\frac{1}{2}\text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})\right). \tag{2}$$

The expectation of the inverse-Wishart distribution is $\boldsymbol{S}/(\nu - k - 1)$.

For all illustrations of the inverse-Wishart distribution in this paper, we assume that $\boldsymbol{S}$ is a $k \times k$ identity matrix, denoted by $\boldsymbol{I}_k$, which makes the variables exchangeable. However, our method can easily be used with any other scale matrix. The distribution is proper if and only if $\nu \geq k$, and the first moment only exists if $\nu > k+1$). Furthermore, if $\boldsymbol{\Sigma}$ follows an inverse-Wishart distribution, then any submatrix of q (possibly permuted) variables is also inverse-Wishart (Eaton 1983): $\boldsymbol{\Sigma}_D \sim \text{inv-Wishart}_{\nu-k+q}(\boldsymbol{S}_D^{-1})$ where $\boldsymbol{S}_D$ is such a submatrix of $\boldsymbol{S}$. If we take $\nu = k + d$ (where d is a positive constant), then the degrees of freedom of the distribution of $\boldsymbol{\Sigma}_D$ are $q + d$ and do not depend on k. The degrees of freedom of the distribution of a submatrix then parallels that of the original matrix. As a consequence, we can focus on the 4×4 covariance matrix distribution to study most (but not all) properties of the inverse-Wishart, because the joint marginal of two correlation coefficients ρ_{ij} and ρ_{km} without a common variable requires at least four dimensions.

2.1. General strategy

In our visualization method we start with sampling L (typically $L = 1000$) covariance matrices $\boldsymbol{\Sigma}$ (or correlation matrices $\boldsymbol{R}$) from their distribution and then plot the sampling distribution of various statistics in four layers or parts. The first layer consists of univariate histograms of the correlations and the logarithm of the standard deviations. The second layer is a set of bivariate scatterplots of variances or correlations. In the third layer, we display the distribution of two variances and a correlation by means of overlaying isodensity contours, and three-dimensional scatterplots of three correlations. Finally, the fourth layer is based on reducing the entire covariance matrix into scalar measures called the effective variance and the effective dependence (Peña and Rodríguez 2003).

Our four-layered graphical representation reveals different aspects of the covariance distribution that can be used to compare the implications of different values of hyperparameters or different distributions. A detailed explanation of the method is given below, but Figure 1 provides an illustration of the method, as generated by our R function `VisCov` (from the package with the same name available on CRAN).

2.2. Layer 1: Histograms of $\log(\sigma_i)$ and ρ_{ij}

For the $L = 1000$ draws, we display the histogram of the logarithm of the standard deviations and the histogram of a correlation ρ_{ij} (typically ρ_{12}). Given the assumption of exchangeability, the histogram for each standard deviation is the same, and the same

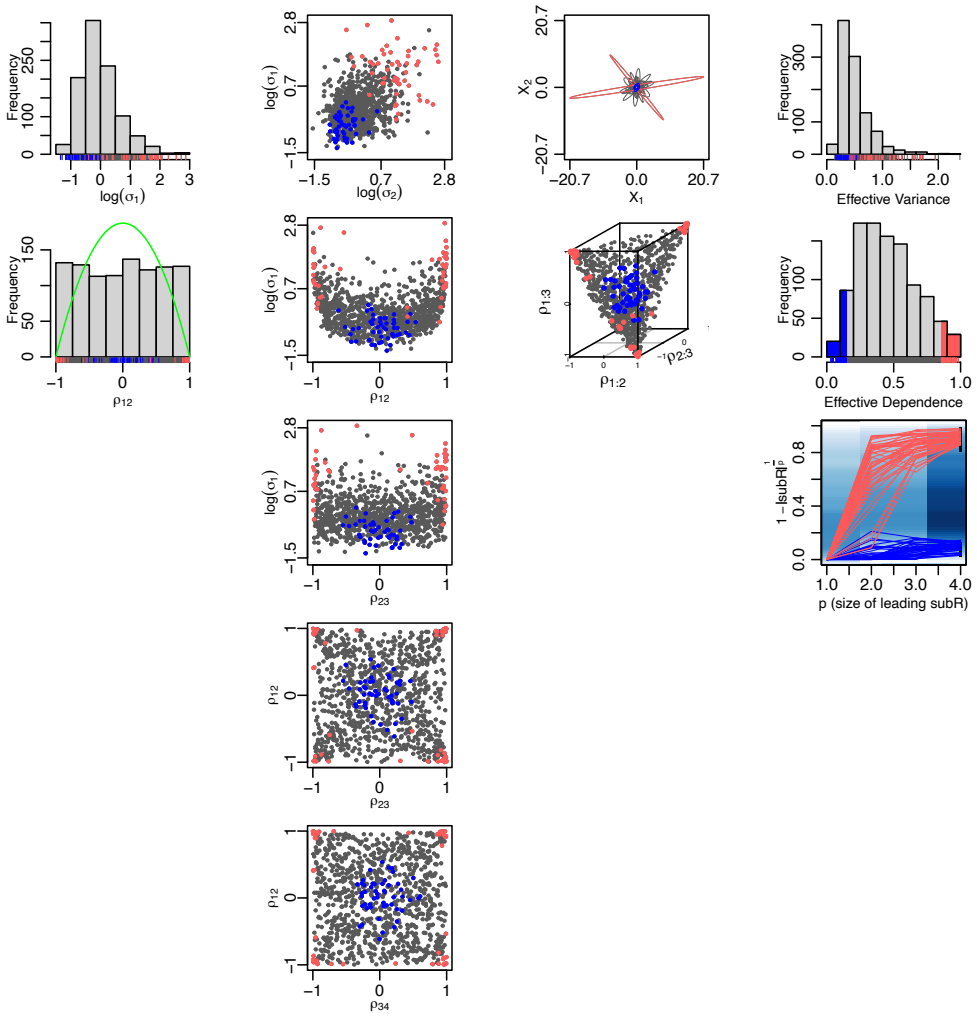


Figure 1: Visualization of an inverse-Wishart distribution with dimension $k = 4$, degrees of freedom $k + 1$ and an identity scale matrix, as generated by the R function `VisCov` from the package with the same name (Example 1 in the example file of `VisCov`). To construct this plot, 1000 covariance matrices were sampled from the inverse-Wishart distribution. Each column of plots refers to a different layer of the visualization method (univariate, bivariate, trivariate and multivariate). The first column shows two histograms (of a log standard deviation and of correlation). A green reference line has been added to the correlation histogram, which is the density of a correlation of a correlation matrix that is uniformly distributed (see below in the text for a detailed explanation of this case). In the second column, various scatterplots are shown. In the third column, the first plot shows 100 50% equiprobability ellipses of a normal distribution centered at the origin (based on a random subsample of the 1000 covariance matrices). In the last column, histograms of the effective variance and the effective dependence are shown along with a plot that displays effective dependence as a function of a growing dimension of the leading principal submatrix. In all plots representations stemming from covariance matrices with extreme effective dependencies are colored blue or red (for small and large effective dependence, respectively). This figure can be generated by `set.seed(1234); VisCov(title.logical = FALSE)`. See the documentation of `VisCov` for more explanation.

holds for the histogram of each correlation; therefore, it is only necessary to display one of each.

2.3. Layer 2: Scatterplots

From the L draws, we can construct $\binom{k(k+1)/2}{2}$ scatterplots for pairs of covariances, but a smaller number of plots will suffice because of the exchangeability assumption. The scatterplots can be found in the second column of Figure 1, where a scatterplot of variable y versus x will be denoted as (x, y).

If $k = 2$, the two plots (σ_1, σ_2) and (σ_1, ρ_{12}) contain all the relevant information. If $k = 3$, there are four non-redundant plots: (σ_1, σ_2), (σ_1, ρ_{12}), (σ_1, ρ_{23}), (ρ_{12}, ρ_{23}). The difference between (σ_1, ρ_{12}) and (σ_1, ρ_{23}) is that there is an overlap of variables in the former. Finally, if $k > 3$, five scatterplots are necessary when exchangeability is assumed: (σ_1, σ_2), (σ_1, ρ_{12}), (σ_1, ρ_{23}), (ρ_{12}, ρ_{23}), (ρ_{12}, ρ_{34}). In the last plot, the correlations have no variable in common, which illustrates our earlier claim that covariance matrices with $k = 4$ are sufficient to reveal most of the interesting information about the inverse-Wishart distribution.

2.4. Layer 3: Contour plot and three-dimensional scatterplots

Here we look at three elements of the covariance matrix simultaneously in two different ways.

Contour plot. In the contour plot approach (see third column, first plot of Figure 1), we focus on the distribution of a specific 2×2 marginal sub-matrix of $\mathbf{\Sigma}$. For instance, we can look at the sub-matrix formed by variables i and j, defined as

$$\begin{pmatrix} \sigma_i^2 & \sigma_i\sigma_j\rho_{ij} \\ \sigma_i\sigma_j\rho_{ij} & \sigma_j^2 \end{pmatrix}. \tag{3}$$

Given the exchangeability assumption, we take $i = 1$ and $j = 2$ and further assume that the two variables are bivariate normal with mean vector zero. Thus, the 50% equiprobability ellipse (the contour plot in which 50% of the bivariate normal density lies) can be plotted (see Johnson and Wichern 2007). It would be straightforward to use another bivariate distribution if desired.

Each ellipse represents an idealized cloud of points in two dimensions and gives information about the orientation and spread of the points along both axes. To avoid clutter, we usually show fewer contour plots; about 100 seem sufficient to visualize the pattern of isodensity contours.

Three-dimensional scatterplot. We also include a three-dimensional scatterplot of three correlations: ρ_{ij}, $\rho_{i,j+1}$ and $\rho_{i+1,j+1}$. The triplet $(\rho_{ij}, \rho_{i,j+1}, \rho_{i+1,j+1})$ corresponds to a 3×3 correlation submatrix from the $k \times k$ correlation matrix. Given the exchangeability assumption, we can take $i = 1$ and $j = 2$.

The two-dimensional scatterplots of the correlations already suggest some of the intricate relations among the correlations under the positive semi-definiteness constraint, and the three-dimensional scatterplot goes a step further. It has been shown by Rousseeuw and Molenberghs (1994) that the support of the distribution of three correlations is a convex volume (called a elliptical tetrahedron). Any cross-section parallel

to one of the two-dimensional coordinate planes forms an ellipse, implying that the support of the conditional distribution of two correlation coefficients given the third one is elliptical. The general pattern in the second plot of the third column of Figure 1 will often occur throughout this paper: For a sufficiently large sample, the convex hull of the points in the scatterplot will approximately coincide with the elliptical tetrahedron, but the way in which the points are distributed across this volume may differ from one distribution to another.

2.5. Layer 4: Effective variance and dependence

Visualization in four or more dimensions is difficult, but visualization in fewer dimensions cannot capture all relevant information in a covariance distribution. Thus, we also analyze scalar statistics that are a function of the entire covariance matrix.

Peña and Rodríguez (2003) have defined the effective variance V_e of a $k \times k$ covariance matrix $\mathbf{\Sigma}$ to be

$$V_e = |\mathbf{\Sigma}|^{1/k}, \tag{4}$$

and the effective dependence as

$$D_e = 1 - |\boldsymbol{R}|^{1/k}, \tag{5}$$

where $\boldsymbol{R}$ is the correlation matrix derived from $\mathbf{\Sigma}$.

These two statistics facilitate comparisons across different values of k or different-sized submatrices of $\mathbf{\Sigma}$. The first two plots of the last column of Figure 1 give the histograms of the effective variance and the effective dependence under the inverse-Wishart distribution. In a third plot, the effective dependence is shown as a function of a growing dimension of the leading principal submatrix. Specifically, the effective dependence of a leading $i \times i$ submatrix is given on the y-axis as a function of i (with $i = 1, \ldots, k$). Each $k \times k$ correlation matrix defines a line (by letting the leading $i \times i$ submatrix grow) and the collection of lines is smoothed and shown as the light blue background. The plot shows how the effective dependence changes with increasing dimension, and how much variation there is across the distribution. In particular, when k is large, it is useful to visualize how the distribution of effective dependence converges.

The most extreme matrices with respect to effective dependence are indicated by the colored tails (blue for low values and red for large values). The matrices with very low or very high effective dependence can be identified in the histograms via the rug, via the same blue and red color scheme in the scatterplots and contour plots, which illustrates how the effective dependence relates to, for instance, the bivariate distribution of the correlations.

3. The visualization method in action

In this section, we apply the four-layered visualization method to various distributions of covariance matrices, draw implications from the plots, and compare distributions. Due to space limitations, we do not present the complete graphical display (such as in

Figure 1), but focus on plots that facilitate comparisons. The full four-layered plot can be recreated using our `VisCov` function.

3.1. The inverse-Wishart distribution

We first reexamine several features of the inverse-Wishart distribution with $k = 4$ dimensions and $\nu = k + 1 = 5$ degrees of freedom from Figure 1. First, the univariate marginal distribution of a correlation is uniform, which is known analytically when $\nu = k + 1$. Second, the scatterplots reveal a strong positive relationship between the (log) standard deviations, which is also reflected in the contour plot (third layer) where ellipses stretching along only one of the main axes are rare. In this distribution, large ellipses tend to be oriented along one of the two principal diagonals. Third, if two variables are extremely correlated, their standard deviations tend to be large, which shows up in the contour plot as well. Fourth, covariance matrices with a large degree of effective dependence (colored in red) tend to have large (log) standard deviations and extreme correlations, which is also apparent in the contour plots where extreme effective dependence coincides with small volume in the metric space. Covariance matrices with little effective dependence tend to have smaller ellipses and (log) standard deviations.

In Figure 2, two inverse-Wishart distributions are compared with $\nu = k + 1$ and $\nu = k + 50$ in the left and right column, respectively, and $k = 4$ in both cases. This figure includes two types of scatterplots (containing log standard deviations and correlations), contour plots, and histograms of effective dependencies. When $\nu = k + 50$, there is less dependence between the log standard deviations and between the correlations, but the marginal distributions of both log standard deviations and correlations become heavily concentrated. A similar pattern can be seen in the contour plots: The length of the major and minor axes decreases, as well as the eccentricity. For a large number of degrees of freedom, the effective dependence becomes very small, which is in line with the fact that the marginal correlations tend to be close to zero.

Both Figures 1 and 2 support the observation by Gelman et al. (2013) that the inverse-Wishart is restrictive as a prior distribution, in part due to the fact that it has only one degrees of freedom parameter, ν. When the degrees of freedom are set to $\nu = k + 1$, the marginal distribution of the correlation is uniform, but the joint distribution of the correlations is far from uniform. There tends to be an abundance of mass in the extreme corners of its support (see, e.g., the starlike pattern in the middle column, last two scatterplots, of Figure 1) and often severe effective dependence. Increasing the degrees of freedom concentrates the distribution around its expectation (here $\boldsymbol{I}_k/(\nu - k - 1)$), which is a strong prior that may not be appropriate in a particular research situation.

We also compare different values of dimensionality, k, which does not lead to qualitatively different conclusions when looking at the first three layers of our visualization method. However, the distribution of effective dependencies is pushed toward the upper bound of one when k increases (and $\nu = k + 1$); see Figure 3 which presents two histograms of effective dependence for the inverse-Wishart distributions with $k = 4$ and $k = 100$. The intuition behind this conjecture is that the average proportion of explained variability of the variables increases with the number of variables, just as a simple R^2 increases in a regression if more predictors are added.

In order to mitigate the restrictiveness of the inverse-Wishart distribution, separation

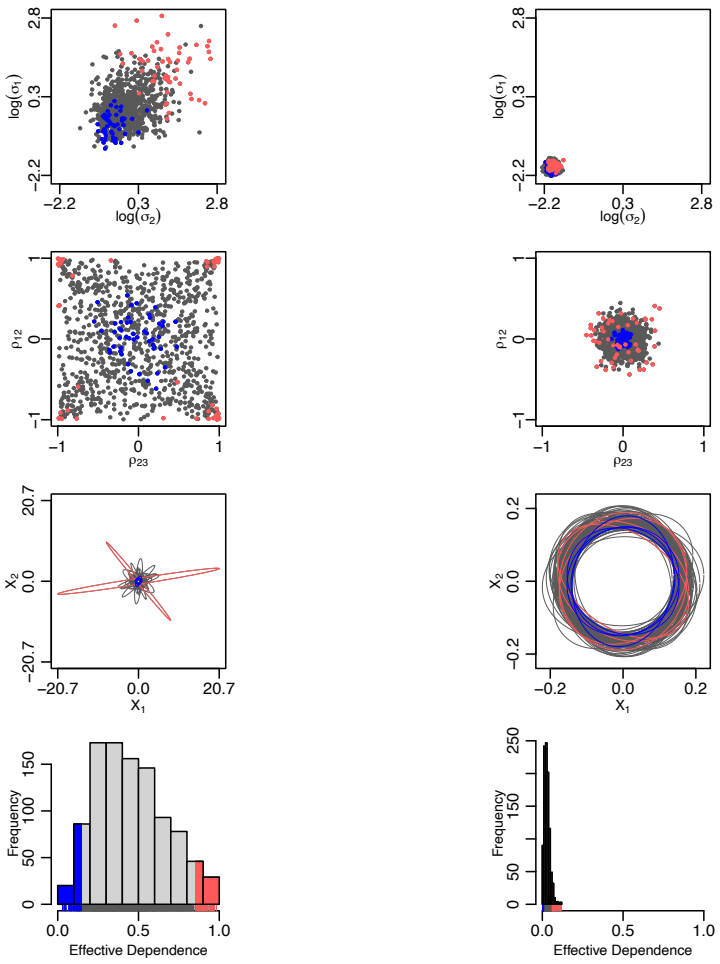


Figure 2: Visualization of inverse-Wishart distributions (using 1000 simulations) with dimension $k = 4$, degrees of freedom $k + 1$ (left column) and $k + 50$ (right column), and an identity scale matrix. The first row represents the scatterplot of two log standard deviations, the second row depicts the scatterplot of two correlations (that share a common variable), the third row shows 100 contour plots for covariance matrics of two variables, and the last row contains histograms of the effective dependencies. The covariance matrices with extreme effective dependencies are colored blue or red (for small and large effective dependence, respectively). The points in the other plots based on these extreme effective dependency matrices are also colored blue and red. The figures shown here can be generated by setting `set.seed(1234)`, followed by running Example 2 of `VisCov`. See the documentation of `VisCov`.

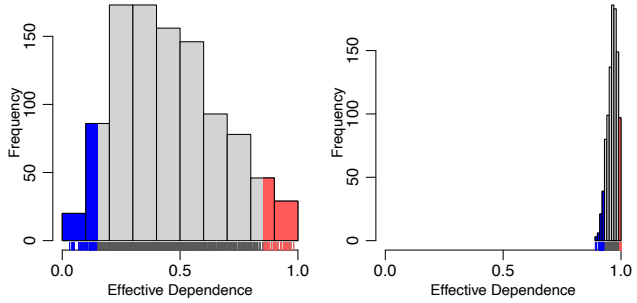


Figure 3: The distribution of the effective dependence of an inverse-Wishart distribution with dimension $k = 4$ (left panel) and $k = 100$ (right panel) based on 1000 simulations with $\nu = k + 1$ and an identity scale matrix. The covariance matrices with extreme effective dependencies are colored blue or red (for small and large effective dependence, respectively). The points in the other plots based on these extreme effective dependency matrices are also colored blue and red. Those figures can be generated by setting `set.seed(1234)`, followed by running Example 3 of `VisCov`. See the documentation of `VisCov`.

strategies have been proposed where the covariance matrix is decomposed so that different priors can be specified on the resulting components. The next three subsections will be devoted to alternatives to the inverse-Wishart implied by three separation strategies based on different decompositions of $\boldsymbol{\Sigma}$.

3.2. Alternative distribution based on separation strategy with marginal correlation matrix from the inverse-Wishart

Barnard et al. (2000) propose the following decomposition:

$$\boldsymbol{\Sigma} = \text{Diag}(\sigma_1, \ldots, \sigma_k) \cdot \boldsymbol{R} \cdot \text{Diag}(\sigma_1, \ldots, \sigma_k), \tag{6}$$

where $\boldsymbol{R}$ has the marginal distribution of the correlation matrix in the inverse-Wishart distribution. The notation $\text{Diag}(\sigma_1, \ldots, \sigma_k)$ refers to a diagonal matrix obtained by placing the σ_i's on the diagonal. In other words, the standard deviations are integrated out, leaving the marginal distribution of the correlation matrix, and then the distribution of the standard deviations can be taken any marginal distribution to form a new joint distribution on the covariance matrix. As derived by Barnard et al. (2000), the kernel of the density function of $\boldsymbol{R}$ is:

$$p(\boldsymbol{R}) \propto |\boldsymbol{R}|^{\frac{(\nu-1)(k-1)}{2}-1} \left(\prod_{i=1}^{k} |\boldsymbol{R}_{ii}| \right)^{-\nu/2}, \tag{7}$$

where $\boldsymbol{R}_{ii}$ is the ith principal sub-matrix of $\boldsymbol{R}$ (obtained from $\boldsymbol{R}$ by removing row and column i). There are several options for the prior distribution on each σ_i (see O'Malley and Zaslavsky 2008). Here, we assume each σ_i is distributed as a standard half-normal:

$$\sigma_i \overset{i.i.d}{\sim} N^+(0, 1), \tag{8}$$

for $i = 1, \ldots, k$.

Thus, the standard deviations are not affected by the degrees of freedom by construction, as is reflected in the first row of Figure 4. Similarly, the correlations are independent of the standard deviations, as is reflected in the second row. However, the univariate and joint distributions of the correlations are similar to those from the inverse-Wishart distribution because they depend on the same degrees of freedom parameter, ν. Increasing ν would move the marginal distribution of a correlation from a uniform distribution toward a peaked distribution around zero. Again, we see a star-like pattern in the bivariate scatterplots. Thus, despite some increased flexibility, this distribution has many of the same problems as the inverse-Wishart distribution.

3.3. Alternative distribution based on separation strategy with uniform prior on the correlation matrix: The LKJ prior

An alternative distribution for $\boldsymbol{\Sigma}$ is obtained with a separation strategy in which the matrix $\boldsymbol{R}$ in Equation (6) has a joint uniform distribution (Barnard et al. 2000) on its support, the space of all $k \times k$ correlation matrices. Here we assume again that each σ_i has a standard half-normal distribution.

Investigating the properties of such distribution requires an efficient algorithm to draw a correlation matrix uniformly, which is not straightforward for $k \geq 3$ due to the positive semi-definiteness restriction. However, Joe (2006) proposed such an algorithm, further refined by Lewandowski et al. (2009), based on the bijective function that exists between the correlation matrix $\boldsymbol{R}$ and $(k-1)(k-2)/2$ partial correlations (viz., correlations between the residuals of two variables when each is regressed on some subset of the other variables). The simplest approach is to condition on all variables before the ith when defining the partial correlation between the ith and jth variables where $i < j$. In that case, each partial correlation can be drawn independently from a symmetric beta distribution spread over the $(-1, 1)$ interval with both shape parameters equal to $\alpha_i = \eta + (k - i - 1)/2$, where $\eta > 0$ is a hyperparameter. Specifically, under such circumstances, Lewandowski et al. (2009) prove that

$$p(\boldsymbol{R}|\eta) = \frac{1}{c(k,\eta)}|\boldsymbol{R}|^{\eta-1}, \tag{9}$$

where $c(k, \eta)$ is a normalization constant, which was first given in Joe (2006); see also Equation (A.2) below. The distribution in Equation (9) is often referred to as the LKJ (Lewandowski-Kurowicka-Joe) distribution (Appendix A in Gelman et al. 2013; Chapter 8 in Lambert 2018). $\boldsymbol{R}$ has a joint uniform distribution if and only if $\eta = 1$; Lewandowski et al. (2009) further prove that the marginal distribution of each correlation is then a symmetric beta distribution over the $(-1, 1)$ interval with both shape parameters $\alpha = k/2$. Hence, the marginal distribution of each correlation becomes more concentrated around zero as k increases in order to satisfy the positive semi-definiteness constraint on the correlations jointly. However, the distribution of $\boldsymbol{\Sigma}$ is not uniform, because its density also depends on the realizations of the standard deviations.

We visualize the implied distribution of $\boldsymbol{\Sigma}$ in Figure 5, with $k = 4$ and $k = 50$. Again, the correlations are independent of the standard deviations by construction. However, unlike the distributions that we have seen thus far, the scatterplot of pairs of correlations

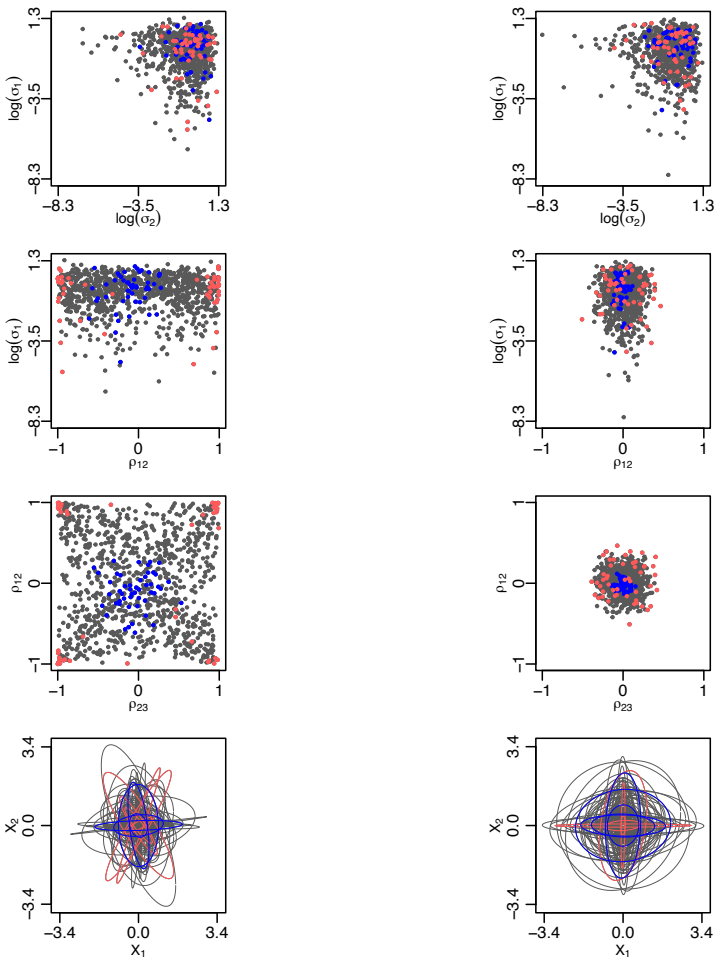


Figure 4: Visualizing a covariance matrix distribution based on decomposing an inverse-Wishart distributed covariance matrix into a correlation matrix and standard deviations. In the left column, $\nu = k+1$ while $\nu = k+50$ in the right column, the dimension is $k = 4$ in both cases, and the standard deviations have a standard half-normal distribution. The number of realizations is 1000 (only 100 50% equiprobability ellipses in the final row are shown). The covariance matrices with extreme effective dependencies are colored blue or red (for small and large effective dependence, respectively). The points in the other plots based on these extreme effective dependency matrices are also colored blue and red. These figures can be generated by setting `set.seed(1234)`, followed by running Example 4 of `VisCov`. See the documentation of `VisCov`.

does not follow a star-like shape. Rather, the correlations seem to be less dependent on each other. Also, a comparison of the evolution of $1 - |\boldsymbol{R}_{p,k}|^{1/p}$ as a function of p reveals two points: First, the range of the effective dependencies is large for small k and very narrow for large k. Second, for large k, the effective dependence of the leading submatrix increases almost linearly, whereas it increases more erratically for small k.

For the asymptotic behavior of the effective dependence, it can be proven that it converges to $1 - \exp(-1) = 0.632$ in probability; see Corollary 1 in Appendix A.

Lastly, we visualize the LKJ distribution on the correlation matrix, setting $k = 100$ and $\eta = 5$ in Equation (9); see Figure 6). It this case, too, the effective dependence converges to $1 - \exp(-1) = 0.632$ in probability for $\eta \in \mathcal{N}$; see Theorem 1 and its proof in Appendix A.

3.4. Alternative overparametrized distribution based on separation strategy: The scaled inverse-Wishart distribution

As a final separation strategy, the scaled inverse-Wishart distribution uses an overparametrized distribution in which there is a tradeoff between sets of parameters. The covariance matrix is now decomposed as follows (see O'Malley and Zaslavsky 2008):

$$\boldsymbol{\Sigma} = \text{Diag}(\xi_1, \ldots, \xi_k) \cdot \boldsymbol{Q} \cdot \text{Diag}(\xi_1, \ldots, \xi_k). \tag{10}$$

$\boldsymbol{Q}$ has an inverse-Wishart distribution with degrees of freedom ν and identity scale matrix:

$$\boldsymbol{Q} \sim \text{inv-Wishart}_{\nu}(\boldsymbol{I}_k). \tag{11}$$

In this case, ξ_i is not a standard deviation because $\boldsymbol{Q}$ does not have 1's on its diagonal. The ith standard deviation is $\xi_i\sqrt{Q_{i,i}}$. Nevertheless, we take the distribution for each ξ_i to be standard half-normal.

The scaling operation has no effect on the correlations, so the correlational properties of the scaled inverse-Wishart distribution are the same as those of the unscaled inverse-Wishart distribution. One motivation for the scaled inverse-Wishart distribution is to mitigate the dependence between the standard deviations and the correlations that plagues the unscaled inverse-Wishart distribution. However, the scaling does not completely eliminate this dependence.

In Figure 7, similar patterns as for the unscaled inverse-Wishart are seen in the scatter and contour plots for $\nu = k + 1$. The dependence between the standard deviation and correlation is weaker, confirming the aforementioned property. When $\nu = k + 50$, the plots differ from those of the unscaled inverse-Wishart, because the distribution of $\boldsymbol{\Sigma}$ is now dominated by $\boldsymbol{\xi}$ (with large variability) as the correlations tend to zero.

3.5. The Wishart distribution

The Wishart distribution, prominent in multivariate statistics (see Johnson and Wichern 2007; Press 1982), is the sampling distribution of $(n-1)\boldsymbol{S}$, where $\boldsymbol{S}$ is the $k \times k$ sample covariance matrix calculated from a sample of size n normal observations on k variables with mean vector $\boldsymbol{\mu}_0$ and population covariance matrix $\boldsymbol{\Sigma}_0$ (see Wishart 1928). As can be seen in Figure 8, the marginal distribution of a covariance matrix

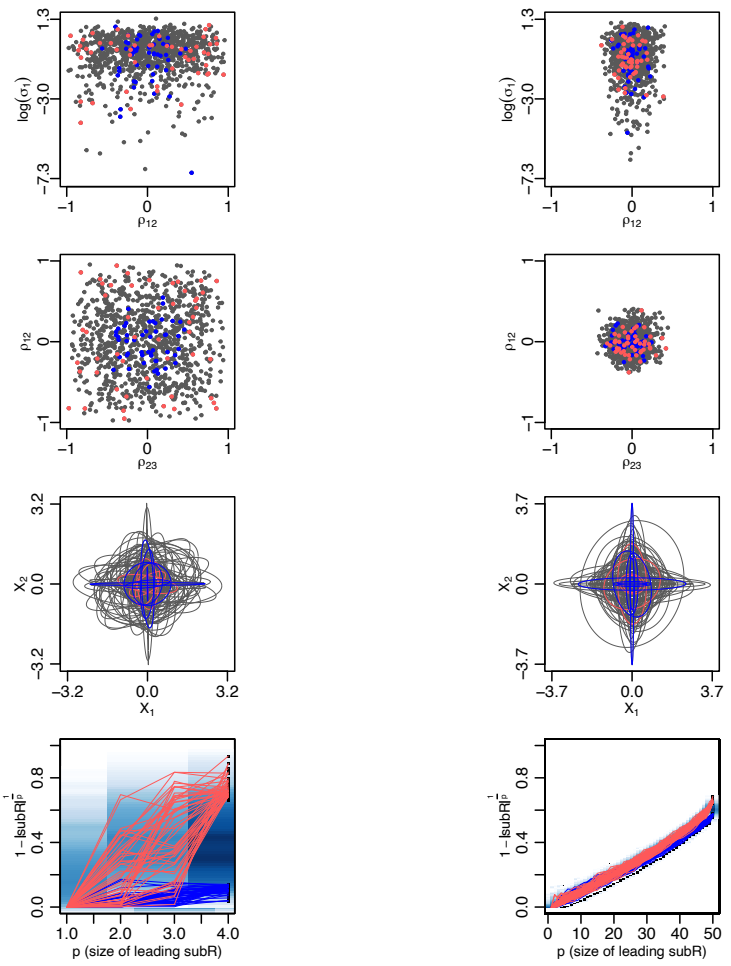


Figure 5: Visualizing a covariance matrix distribution based on a separation strategy with a uniformly distributed correlation matrix and standard half-normal standard deviations. In the left column, $k = 4$ and in the right column $k = 50$. The number of realizations is 1000 (only 100 of the 50% equiprobability ellipses are shown). The covariance matrices with extreme effective dependencies are colored blue or red (for small and large effective dependence, respectively). The points in the other plots based on these extreme effective dependency matrices are also colored blue and red. These figures can be generated by setting `set.seed(1234)`, followed by running Example 5 of `VisCov`. See the documentation of `VisCov`.

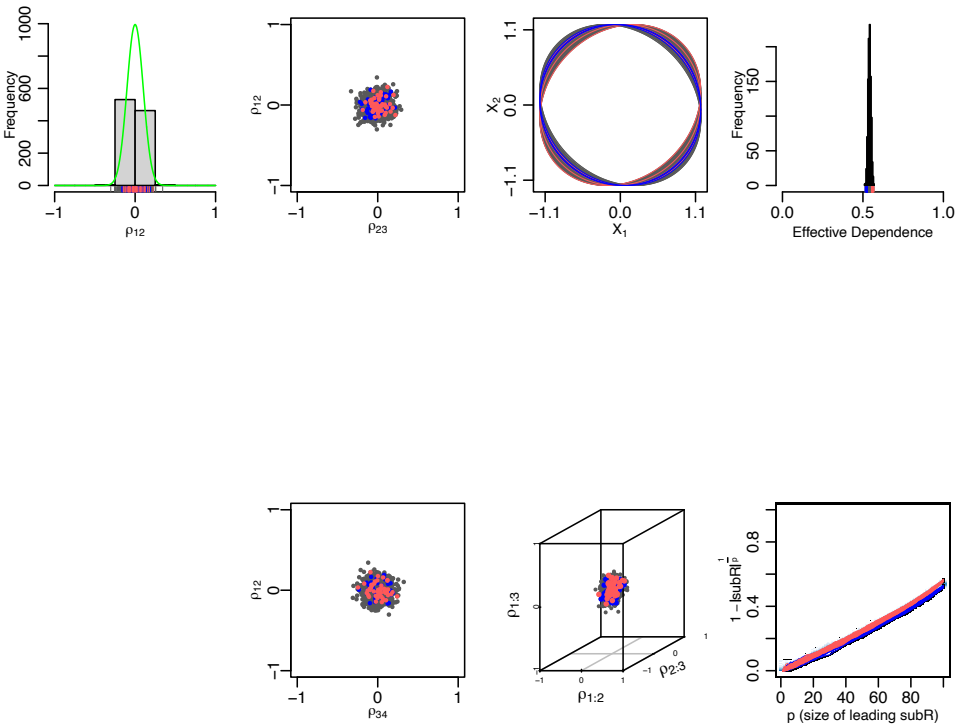


Figure 6: LKJ distribution of the correlation matrix with $k = 100$ and $\eta = 5$ in Equation (9). We present the relevant panels only, while removing those related to variances. The number of realizations is 1000 (only 100 for the 50% equiprobability ellipses are shown). The covariance matrices with extreme effective dependencies are colored blue or red (for small and large effective dependence, respectively). The points in the other plots based on these extreme effective dependency matrices are also colored blue and red. These figures can be generated by setting `set.seed(1234)`, followed by running Example 6 of `VisCov`. See the documentation of `VisCov`.

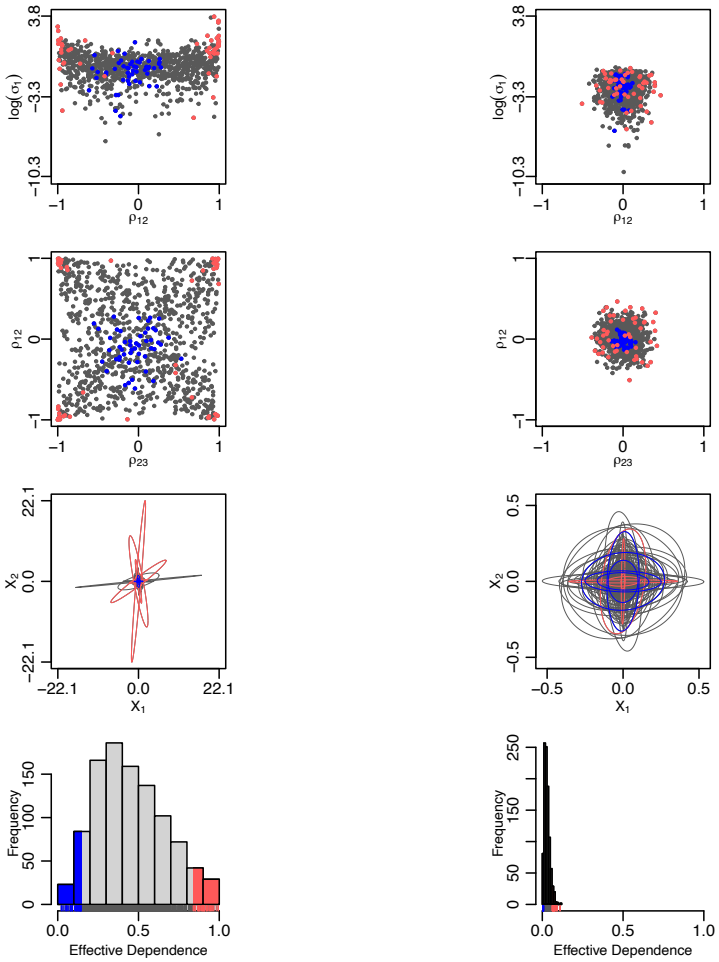


Figure 7: The scaled inverse-Wishart with $\nu = k + 1$ (left column) and $\nu = k + 50$ (right column) with $k = 4$. The number of realizations is 1000 (only 100 50% equiprobability ellipses are shown in the third row). The covariance matrices with extreme effective dependencies are colored blue or red (for small and large effective dependence, respectively). The points in the other plots based on these extreme effective dependency matrices are also colored blue and red. These figures can be generated by setting `set.seed(1234)`, followed by running Example 7 of `VisCov`. See the documentation of `VisCov`.

depends on k (when $\nu = k + 1$). As the dimension increases, there is less variability in standard deviations, correlations and effective dependence. If one seeks similar marginal distributions for the standard deviations when k varies, different values of ν must be used (and not always $\nu = k + 1$).

For the asymptotic behavior of the effective dependence, we have the same results as in the case of the uniform distribution: It can be proven that the effective dependence converges to $1 - \exp(-1) = 0.632$ in probability (see Theorem 2 and its proof in Appendix A).

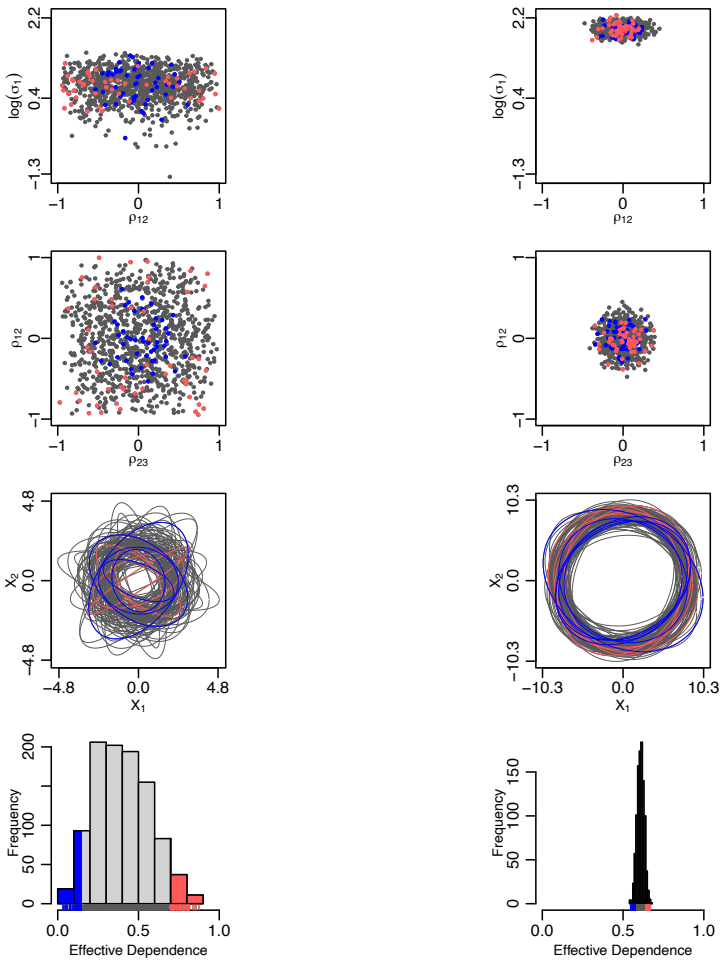


Figure 8: Wishart distribution with $\nu = k + 1$, where $k = 4$ (left column) and $k = 50$ (right column). The number of realizations is 1000 (only 100 50% equiprobability ellipses in the third row are shown). The covariance matrices with extreme effective dependencies are colored blue or red (for small and large effective dependence, respectively). The points in the other plots based on these extreme effective dependency matrices are also colored blue and red. These figures can be generated by setting `set.seed(1234)`, followed by running Example 2 of `VisCov`. See the documentation of `VisCov`.

3.6. A new distribution

Our visualization tool may help in the evaluation of customized distributions. For example, assume that we define the following prior distribution on covariance matrices:

$$\boldsymbol{\Sigma} = \text{Diag}(\sigma_1, \ldots, \sigma_k) \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{D} \cdot \boldsymbol{\Lambda}' \cdot \text{Diag}(\sigma_1, \ldots, \sigma_k), \tag{12}$$

where $\boldsymbol{\Lambda}$ is a $k \times k$ randomly generated orthogonal matrix (draw a $k \times k$ matrix $\boldsymbol{W}$ with standard normal deviates, calculate the singular value decomposition $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$ and then set $\boldsymbol{\Lambda} = \boldsymbol{U}\boldsymbol{V}'$), $\boldsymbol{D}$ is a diagonal matrix of eigenvalues, the marginal distribution of which can be any distribution with positive support. Here we draw each diagonal element of $\boldsymbol{D}$ from a beta$(0.5, 5)$ distribution, which tends to yield a few eigenvalues close to one and many eigenvalues close to zero. Again, each σ_i is a standard deviation and has a standard half-normal distribution.

We investigate the properties of this customized distribution in Figure 9. The correlations are concentrated around zero, but not as strongly as we have seen with other covariance distributions. The effective dependence is a function of the eigenvalues only, and is centered around the mean value 0.67 (a value close to the limiting value for the uniformly distributed correlation matrices as k goes to infinity).

4. Comparison of distributions

In order to select a prior in an empirical application, researchers must know the properties of the various choices for a covariance distribution. In this section, we compare the four distributions discussed above with regard to the marginal and joint distributions of the covariances and the dependence on k.

First, for the inverse-Wishart distribution, as ν gets larger, the correlations are concentrated around zero. For the scaled inverse-Wishart, a similar pattern is obvious for the correlation when $\nu = k+1$, but the standard deviations depend heavily on the prior for each ξ_i. In the previous section, we drew these scale parameters independently from a standard half-normal distribution, so the dependence among the standard deviations is small and driven by $\boldsymbol{Q}$. When $\boldsymbol{R}$ is the marginal correlation matrix of the inverse-Wishart distributed covariance matrix, the correlations are more dependent than when $\boldsymbol{R}$ is given a joint uniform distribution.

Second, the size of the covariance matrix affects some of its properties in different ways across distributions. For the distributions derived from the inverse-Wishart, we can invoke an exchangeability property; hence, only a few plots are needed to understand univariate and bivariate properties of the covariances, even if k is large. However, the effective dependence is a function of k and its sampling distribution depends on the distribution of the covariance matrix. In particular, if $\boldsymbol{R}$ is given a joint uniform distribution, then the effective dependence is bounded away from one as k goes to infinity, which is not the case for the inverse-Wishart distribution. If a researcher wants to use a prior that is dependent on the dimensionality, our visualization tool can be used to gauge its properties.

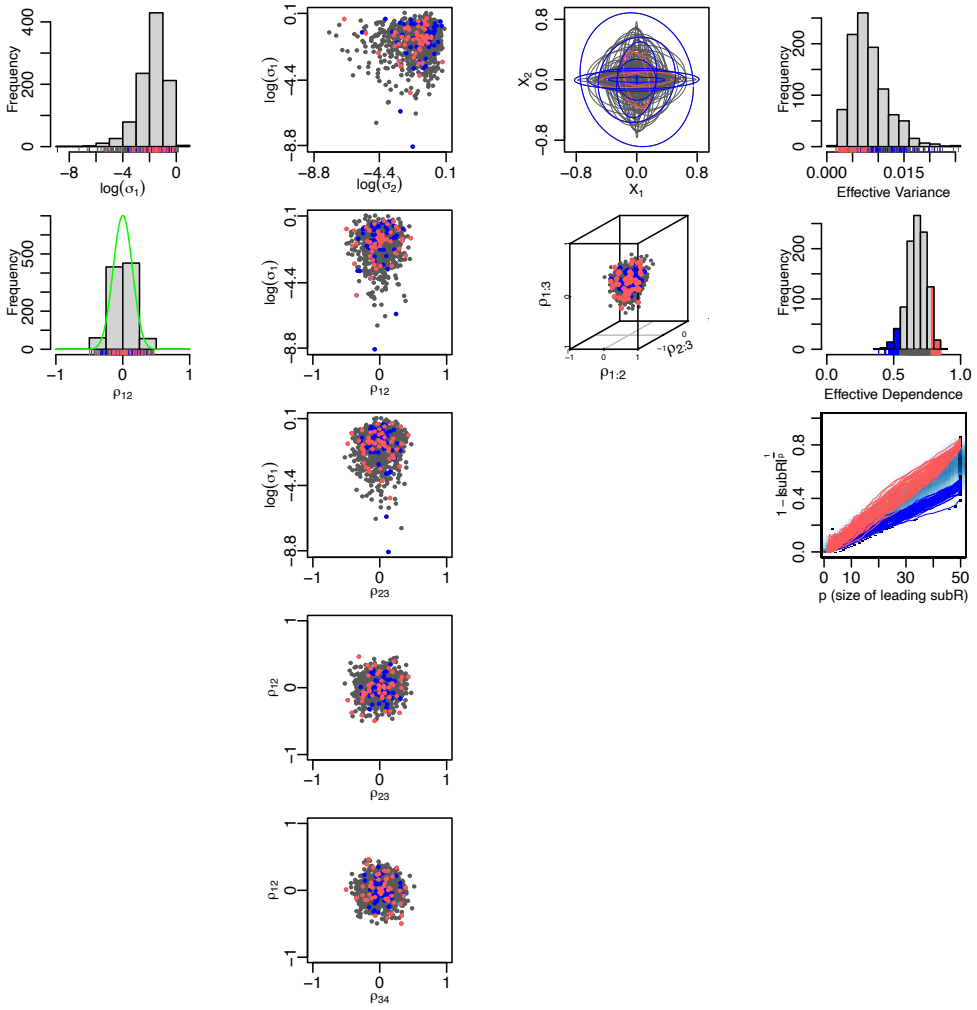


Figure 9: Visualization of a customized distribution (see Equation (3.6) and the text below) of covariance matrices. To construct the plot, 1000 covariance matrices (dimension $k = 50$) were simulated and plotted. The covariance matrices with extreme effective dependencies (50 instances) are colored blue or red (for small and large effective dependence, respectively). The points in the other plots based on these extreme effective dependency matrices are also colored blue and red. These figures can be generated by setting `set.seed(1234)`, followed by running Example 9 of `VisCov`. See the documentation of `VisCov`.

5. An empirical example

We demonstrate the use of our tools to visualize the prior and posterior covariance distributions for a model that was fit to data from the Survey of Consumer Expectations, collected by the Federal Reserve Bank of New York (FRBNY). On a monthly basis, FRBNY asks a sample of respondents to provide point predictions of annual inflation, 12 months, 36 months, and 60 months into the future (i.e., $k = 3$). Here we analyze the median point predictions from January 2023 until January 2025 (25 months of data). The median point predictions are centered at zero. The observed covariance matrix of the $n = 25$ observations is

$$\boldsymbol{S} = \begin{pmatrix} 0.690 & 0.101 & -0.016 \\ 0.101 & 0.067 & 0.003 \\ -0.016 & 0.003 & 0.013 \end{pmatrix}. \tag{13}$$

To model these data, we assume that every (median) point prediction triplet $y_i = (y_{i1}, y_{i2}, y_{i3})^T$ is a draw from a trivariate normal model with mean zero and an unstructured covariance matrix:

$$y_i \sim N\left(0, \boldsymbol{\Sigma}\right),$$

with 0 being the trivariate zero vector.

Four different Bayesian analyses will be conducted. In the first two, we will use inverse-Wishart priors for Σ:

$$\boldsymbol{\Sigma} \sim \text{inv-Wishart}_{k+1=4}(\boldsymbol{I}_{k=3})$$

and

$$\boldsymbol{\Sigma} \sim \text{inv-Wishart}_{k+10=13}(\boldsymbol{I}_{k=3}).$$

In the next two analyses, we will use a decomposition of the covariance matrix:

$$\boldsymbol{\Sigma} = \text{Diag}(\xi_1, \xi_2, \xi_3) \cdot \boldsymbol{R} \cdot \text{Diag}(\xi_1, \xi_2, \xi_3),$$

where $\boldsymbol{R}$ is the correlation matrix and ξ_j is the standard deviation of the median point prediction j ($j = 1, 2, 3$), with the standard half-normal prior for ξ_j and one of the following two LKJ priors for the correlation matrix:

$$\boldsymbol{R} \sim \text{LKJ}(\eta = 1) \text{ or } \boldsymbol{R} \sim \text{LKJ}(\eta = 5).$$

We run the analyses using Stan (Stan Development Team 2024a,b). All $\widehat{R}$'s were smaller than 1.001. The posterior means and standard deviations are shown in Table 1.

Next, we visualize the prior and posterior distributions for $\boldsymbol{\Sigma}$ using `VisCov`. In our visualization we focus on the marginal distribution of ρ_{12}, the joint distribution of ρ_{12} and $\log(\sigma_1)$, the joint distribution of ρ_{12} and ρ_{23}, and a sample of 50% equiprobability ellipses. The results can be found in Figure 10. From these comparisons within each figure and across figures, several observations can be made.

First, when using Bayes' rule for inferential statistics, we are reassigning probability mass from the prior in light of the observed data. By comparing the left and right

		IW4	IW13	LKJ1	LKJ5
σ_1^2	mean	0.707	0.517	0.763	0.730
	sd	0.207	0.129	0.234	0.220
σ_{12}	mean	0.098	0.071	0.101	0.077
	sd	0.060	0.037	0.055	0.043
σ_{13}	mean	−0.016	−0.011	−0.017	−0.013
	sd	0.041	0.025	0.023	0.019
σ_2^2	mean	0.105	0.077	0.077	0.073
	sd	0.031	0.019	0.024	0.022
σ_{23}	mean	0.003	0.002	0.004	0.003
	sd	0.015	0.010	0.008	0.006
σ_3^2	mean	0.053	0.039	0.016	0.015
	sd	0.015	0.010	0.005	0.005

Table 1: Posterior means and standard deviations for the unique parameters of $\mathbf{\Sigma}$ using four different priors. See text for more details.

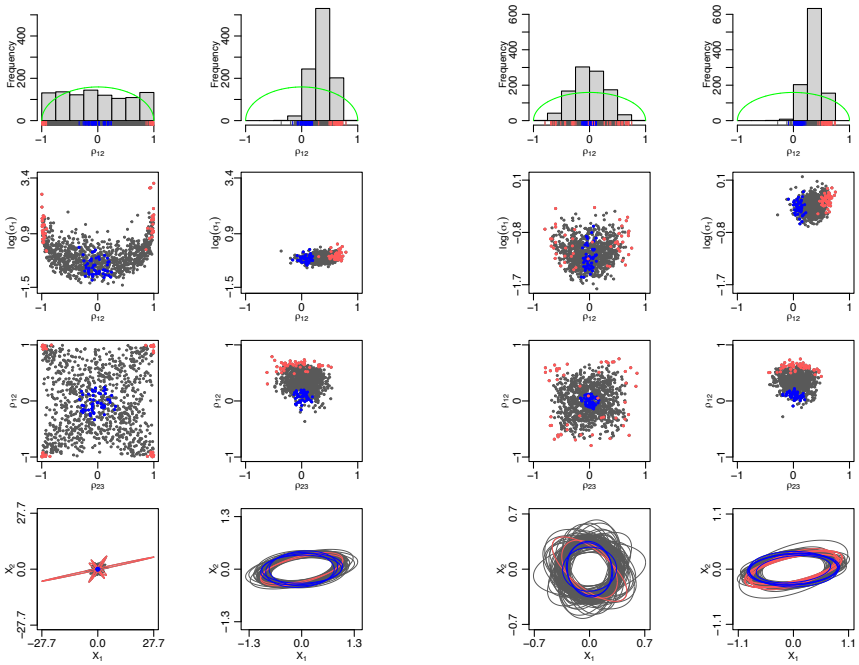
column in each figure, this reassignment of probability from prior to posterior for a distribution on covariance matrices can be seen in action. Second, the more informative a prior is, the more the posterior mass is concentrated in a particular region in the space. Comparing Panel (a) to Panel (b) and Panel (c) to Panel (d), this can be seen clearly.

6. Conclusion

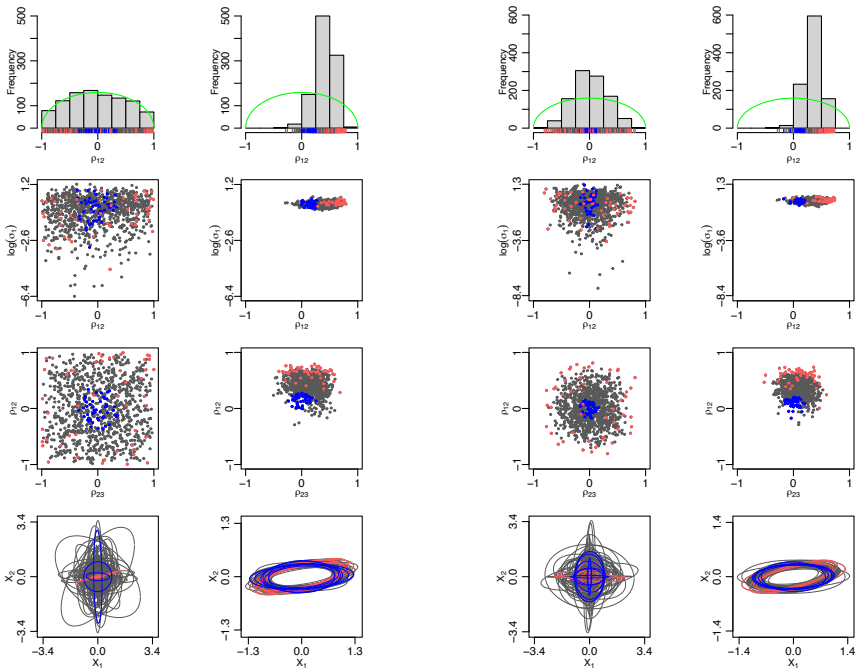
We have introduced a four-layered visualization method for a distribution of covariance matrices that compromises histograms, scatterplots and contour plots. The four layers of plots complement each other, enabling a researcher to visualize distributions of covariance matrices from different perspectives. As we have seen in the examples, this novel visualization method effectively reveals properties of distributions that were not easy to understand analytically.

We take advantage of the exchangeability (i.e., invariance under permutations of variables) in many standard prior distributions of covariance matrices to efficiently and effectively display these highly multivariate distributions using a tableau of low-dimensional displays. Our method can be applied to any proper distribution for covariance matrices that satisfies this exchangeability property. Hence, the method may be useful not only to deepen our understanding of existing distributions, but also to better understand newly proposed distributional families. Unlike typical uses of statistical graphics for data exploration, here we are using visualization to better understand complex mathematical distributions.

An obvious limitation of the method is that it cannot visualize at once a family of distributions that is dependent on the dimensionality k of the variable set; in that case, it is necessary to draw plots for at least several values of k.



(a) inv-Wishart($\nu = 4$) (b) inv-Wishart($\nu = 13$)



(c) LKJ($\eta = 1$) (d) LKJ($\eta = 5$)

Figure 10: Comparison of prior and posterior for the models with four different priors in the text (dimension $k = 3$). Each plot is based on 1000 draws (except the concentration ellipses, which only use 200 draws). In each plot, the left column is the visualization of the prior and the right column is the visualization of the posterior.

7. Code and data availability

The code and data to generate Figures 1–9 in the present paper are available in the R package `VisCov` from the CRAN repository with dependencies on `clusterGeneration`, `bayesm`, `MASS`, `TeachingDemos`, `scatterplot3d` and `KernSmooth`. The code and data for the empirical example are at KU Leuven GitLab: https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0019524/empirical-application-viscov-paper/.

Acknowledgments

We thank the U.S. Institute of Education Science grants #ED-GRANTS-032309-005 and #R305D090006-09A, the U.S. National Science Foundation grant #SES-1023189, the U.S. Department of Energy grant #DE-SC0002099, the U.S. National Security Agency grant #H98230-10-1-0184, the Research Fund of the University of Leuven grants GOA/10/02 and C14/23/062, and the Belgian Government grant IAP P6/03 for partial support of this work. Further, we thank the MEXT Project for Seismology Toward Research Innovation with Data of Earthquake (STAR-E) Grant JPJ010217, and Grant-in-Aid for Fund for the Promotion of Joint International Research (International Collaborative Research) No. 23KK0181. We also thank Kristof Meers for his help in updating the VisCov package.

The data used in the empirical application come from the Survey of Consumer Expectations (SCE) by Federal Reserve Bank of New York (FRBNY). The SCE data are available without charge at http://www.newyorkfed.org/microeconomics/sce and may be used subject to license terms posted there. FRBNY disclaims any responsibility for this analysis and interpretation of the SCE data.

References

Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. New York, Dover Publications.

Alvarez, I., Niemi, J., and Simpson, M. (2014). Bayesian inference for a covariance matrix. *arXiv preprint arXiv:1408.4050*.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. New York, Wiley.

Barnard, J., McCulloch, R., and Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10:1281–1311.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. New York, Wiley.

Cook, D. and Swayne, D. F. (2007). *Interactive and Dynamic Graphics for Data Analysis with R and GGobi*. New York, Springer.

Daniels, M. J. and Kass, M. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94:1254–1263.

De Boeck, P., DeKay, M. L., and Pek, J. (2024). Adventitious error and its implications for testing relations between variables and for composite measurement outcomes. *Psychometrika*, 89(3):1055–1073.

Eaton, M. L. (1983). *Multivariate Statistics: A Vector Space Approach*. New York, Wiley.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *American Statistician*, 56:316–324.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, third edition*. London, CRC Press.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

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

Johnson, R. and Wichern, D. (2007). *Applied Multivariate Statistical Analysis*. Saddle River, N.J., Prentice Hall.

Lambert, B. (2018). *A Student's Guide to Bayesian Statistics*. Sage Publications Ltd.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100:1989–2001.

Liu, H., Zhang, Z., and Grimm, K. J. (2016). Comparison of inverse Wishart and separation-strategy priors for Bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(3):354–367.

Merkle, E. C., Ariyo, O., Winter, S. D., and Garnier-Villarreal, M. (2023). Opaque prior distributions in Bayesian latent variable models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 19:1055–1073.

O'Malley, A. and Zaslavsky, A. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103:1405–1418.

Oravecz, Z., Tuerlinckx, F., and Vandekerckhove, J. (2009). A hierarchical Ornstein-Uhlenbeck model for continuous repeated measurement data. *Psychometrika*, 74:395–418.

Owen, J. and Rabinovitch, R. (1983). On the class of elliptical distributions and their applications to the theory of portfolio choice. *Journal of Finance*, 38:745–752.

Peña, D. and Rodríguez, J. (2003). Descriptive measures of multivariate scatter and linear dependence. *Journal of Multivariate Analysis*, 85:361–374.

Peters, G. W., Dong, A. X., and Kohn, R. (2014). A copula based Bayesian approach for paid–incurred claims models for non-life insurance reserving. *Insurance: Mathematics and Economics*, 59:258–278.

Press, S. J. (1982). *Applied Multivariate Analysis Using Bayesian and Frequentist Methods of Inference*. Malabar, Fla., Krieger.

Rousseeuw, P. and Molenberghs, G. (1994). The shape of correlation matrices. *American Statistician*, 48:276–279.

Stan Development Team (2024a). Rstan: The R interface to Stan. R package version 2.x, https://mc-stan.org/rstan.

Stan Development Team (2024b). *Stan Modeling Language Users Guide and Reference Manual*. https://mc-stan.org.

Theus, M. and Urbanek, S. (2008). *Interactive Graphics for Data Analysis: Principles and Examples*. London, CRC Press.

Valero-Mora, P., Young, F., and Friendly, M. (2003). Visualizing categorical data in ViSta. *Computational Statistics and Data Analysis*, 43:495–508.

Von Rosen, D. (1988). Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics*, 15:97–109.

Wishart, J. (1928). The generalized product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A:32–52.

Yang, R. and Berger, J. (1994). Estimation of a covariance matrix using the reference prior. *Annals of Statistics*, 22:1195–1211.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York, Wiley.

A. Asymptotic behavior of effective dependence

We provide two mathematical results of asymptotic behavior of effective dependence for LKJ distribution and Wishart distribution.

Theorem 1. *Let $\boldsymbol{R}$ be a k-dimensional positive definite correlation matrix. If $\boldsymbol{R}$ follows a LKJ distribution, then, for $\forall \eta \in (0, \infty)$ in Equation (9),*

$$1 - |\boldsymbol{R}|^{1/k} \xrightarrow{P} 1 - \exp(-1) \quad \textit{when } k \to \infty.$$

Proof. By definition:

$$\begin{aligned} E(|\boldsymbol{R}|^{\frac{1}{k}}) &= \int |\boldsymbol{R}|^{\frac{1}{k}} f(\boldsymbol{R}) d\boldsymbol{R} \\ &= \frac{1}{c(k,\eta)} \int |\boldsymbol{R}|^{\frac{1}{k}+\eta-1} d\boldsymbol{R} \\ &= \frac{c(k, \eta + 1/k)}{c(k,\eta)}, \end{aligned} \tag{A.1}$$

where $c(k,\eta)$ and $c(k, \eta + 1/k)$ are normalizing constants defined in Equation (9). To evaluate this expression, we use the results on the normalizing constant in Equation (9) by Joe (2006) (we adapt the notation for the present paper):

$$c(k,\eta) = 2^{\sum_{i=1}^{k-1}(2\eta-2+i)i} \prod_{i=1}^{k-1} \text{Beta}\Big(\eta + \frac{i-1}{2}, \eta + \frac{i-1}{2}\Big)^i. \tag{A.2}$$

By applying this formula to (A.1), $E(|\boldsymbol{R}|^{\frac{1}{k}})$ becomes:

$$E(|\boldsymbol{R}|^{\frac{1}{k}}) = \frac{2^{\sum_{i=1}^{k-1}(2\eta+\frac{2}{k}-2+i)i} \prod_{i=1}^{k-1} \{\text{Beta}(\frac{1}{2}(i-1)+\eta+\frac{1}{k}, \frac{1}{2}(i-1)+\eta+\frac{1}{k})\}^i}{2^{\sum_{i=1}^{k-1}(2\eta-2+i)i} \prod_{i=1}^{k-1} \{\text{Beta}(\frac{1}{2}(i-1)+\eta, \frac{1}{2}(i-1)+\eta)\}^i}. \tag{A.3}$$

Next, we apply the mean value theorem to the logarithm of $E(|\boldsymbol{R}|^{\frac{1}{k}})$ (a prime $'$ refers to a first derivative and logBeta$(\cdot)$ is the logarithm of the beta function):

$$\begin{aligned} \log(E(|\boldsymbol{R}|^{\frac{1}{k}})) &= (k-1)\log 2 + \sum_{i=1}^{k-1} i \Big\{ \text{logBeta}\left(\frac{1}{2}(i-1)+\eta+\frac{1}{k}, \frac{1}{2}(i-1)+\eta+\frac{1}{k}\right) - \\ &\qquad \text{logBeta}\left(\frac{1}{2}(i-1)+\eta, \frac{1}{2}(i-1)+\eta\right)\Big\} \\ &= (k-1)\log 2 + \sum_{i=1}^{k-1} \frac{i}{k} \text{logBeta}'\left(\frac{1}{2}(i-1)+\eta+s_i, \frac{1}{2}(i-1)+\eta+s_i\right) \\ &= (k-1)\log 2 + \sum_{i=1}^{k-1} \frac{i}{k} \frac{\text{Beta}'\left(\frac{1}{2}(i-1)+\eta+s_i, \frac{1}{2}(i-1)+\eta+s_i\right)}{\text{Beta}\left(\frac{1}{2}(i-1)+\eta+s_i, \frac{1}{2}(i-1)+\eta+s_i\right)} \\ &= (k-1)\log 2 + \frac{2}{k} \sum_{i=1}^{k-1} i \Big\{ \psi\left(\frac{1}{2}(i-1)+\eta+s_i\right) - \psi(i-1+2\eta+2s_i)\Big\}, \end{aligned} \tag{A.4}$$

where $\exists s_i \in (0, 1/k)$ by the mean value theorem and $\psi(\cdot)$ is the digamma function, defined as $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. Note that $d \log \text{Beta}(x, x)/dx = 2\psi(x) - 2\psi(2x)$.
As $\psi(x) = \log x - \frac{1}{2x} + O(\frac{1}{x^2})$ (Abramowitz and Stegun 1972), we find that:

$$\begin{aligned}
\log(E(|\boldsymbol{R}|^{\frac{1}{k}})) &= (k-1)\log 2 + \frac{2}{k}\sum_{i=1}^{k-1} i\Bigg\{ \log\left(\frac{1}{2}(i-1) + \eta + s_i\right) - \frac{1}{i-1+2\eta+2s_i} \\
&\qquad - \log(i-1+2\eta+2s_i) + \frac{1}{2(i-1+2\eta+2s_i)} + O\left(\frac{1}{i^2}\right)\Bigg\} \\
&= (k-1)\log 2 + \frac{2}{k}\sum_{i=1}^{k-1} i\Big\{ -\log 2 - \frac{1}{2(i-1+2\eta+2s_i)} + O\left(\frac{1}{i^2}\right)\Big\} \\
&= \frac{2}{k}\sum_{i=1}^{k-1}\Big\{ -\frac{i}{2(i-1+2\eta+2s_i)} + O\left(\frac{1}{i}\right)\Big\} \\
&= \frac{2}{k}\sum_{i=1}^{k-1}\Big\{ -\frac{1}{2} + \frac{-1+2\eta+2s_i}{2(i-1+2\eta+2s_i)} + O\left(\frac{1}{i}\right)\Big\} \\
&= -\frac{k-1}{k} + \frac{2}{k}\sum_{i=1}^{k-1}\Big\{\frac{-1+2\eta+2s_i}{2(i-1+2\eta+2s_i)} + O\left(\frac{1}{i}\right)\Big\} \\
&= -\frac{k-1}{k} + \frac{1}{k}\sum_{i=1}^{k-1} O\left(\frac{1}{i}\right). \qquad \text{(A.5)}
\end{aligned}$$

Consequently, this leads to:

$$\lim_{k\to\infty} \log(E(|\boldsymbol{R}|^{\frac{1}{k}})) = -1, \qquad \text{(A.6)}$$

and also:

$$\lim_{k\to\infty} E(|\boldsymbol{R}|^{\frac{1}{k}}) = \exp(-1). \qquad \text{(A.7)}$$

In the same way, it can be shown that $\lim_{k\to\infty} E(|\boldsymbol{R}|^{\frac{2}{k}}) = \exp(-2)$. Thus,

$$\lim_{k\to\infty} \text{Var}(|\boldsymbol{R}|^{\frac{1}{k}}) = \exp(-2) - (\exp(-1))^2 = 0. \qquad \text{(A.8)}$$

These results imply that the effective dependence converges to $1 - \exp(-1) = 0.632$ in probability. □

Corollary 1. *Let $\boldsymbol{R}$ be a k-dimensional positive definite correlation matrix. If $\boldsymbol{R}$ follows a uniform distribution, then,*

$$1 - |\boldsymbol{R}|^{1/k} \xrightarrow{P} 1 - \exp(-1) \quad \textit{when } k \to \infty.$$

Proof. The uniform distribution is a special case of LKJ distribution with $\eta = 1$ in Equation (9). Hence the proposition is trivial from Theorem 1. □

Theorem 2. *Let $\boldsymbol{R}$ be a k-dimensional positive definite correlation matrix. If $\boldsymbol{R}$ follows the marginal distribution of the correlation matrix derived from $\boldsymbol{\Sigma}$ that has a Wishart distribution with $\nu = k + d$ ($d \geq 1$) degrees of freedom and scale matrix $\boldsymbol{I}_k$, and d is irrespective of k, then,*

$$1 - |\boldsymbol{R}|^{1/k} \xrightarrow{P} 1 - \exp(-1) \quad \textit{when } k \to \infty.$$

Proof. We need to consider a transformation of variables from $\boldsymbol{\Sigma}$ to the correlation matrix $\boldsymbol{R}$ and the standard deviations $\boldsymbol{S} = \text{Diag}(\xi_1, \ldots, \xi_k)$. Since the Jacobian of this transformation is given by $2^k(\prod_{i=1}^k \xi_i)^k$ (Barnard et al. 2000),

$$\begin{aligned} f(\boldsymbol{R}, \boldsymbol{S}) &= f(\boldsymbol{\Sigma}) \times \text{Jacobian} \\ &= c_2^{-1}|\boldsymbol{\Sigma}|^{\frac{\nu-k-1}{2}} \exp\left(-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma})\right) 2^k \left(\prod_{i=1}^k \xi_i\right)^k \\ &= 2^k c_2^{-1}|\boldsymbol{R}|^{\frac{\nu-k-1}{2}} \prod_{i=1}^k \xi_i^{\nu-1} \exp\left(-\frac{1}{2}\xi_i^2\right), \end{aligned} \tag{A.9}$$

where c_2 is the normalizing constant for the Wishart. The density function of $\boldsymbol{R}$ is then:

$$\begin{aligned} f(\boldsymbol{R}) &= \int_0^\infty f(\boldsymbol{R}, \boldsymbol{S}) d\xi_1 \ldots d\xi_k \\ &\propto |\boldsymbol{R}|^{\frac{\nu-k-1}{2}}. \end{aligned} \tag{A.10}$$

Because of Equation (9), we have:

$$f(\boldsymbol{R}) = \frac{1}{c(k, (\nu-k+1)/2)} |\boldsymbol{R}|^{\frac{\nu-k-1}{2}}. \tag{A.11}$$

This leads to:

$$\begin{aligned} E(|\boldsymbol{R}|^{1/k}) &= \int |\boldsymbol{R}|^{\frac{1}{k}} f(\boldsymbol{R}) d\boldsymbol{R} \\ &= \frac{1}{c(k, (\nu-k+1)/2)} \int |\boldsymbol{R}|^{\frac{\nu-k-1}{2}+\frac{1}{k}} d\boldsymbol{R} \\ &= \frac{c(k, (\nu-k+1)/2 + 1/k)}{c(k, (\nu-k+1)/2)}. \end{aligned} \tag{A.12}$$

From Equation (refJoe):

$$E(|\boldsymbol{R}|^{\frac{1}{k}}) = \frac{2^{\sum_{i=1}^{k-1}(d-1+\frac{2}{k}+i)i} \prod_{i=1}^{k-1} \left\{\text{Beta}(\frac{1}{2}(i+d)+\frac{1}{k}, \frac{1}{2}(i+d)+\frac{1}{k})\right\}^i}{2^{\sum_{i=1}^{k-1}(d-1+i)i} \prod_{i=1}^{k-1} \left\{\text{Beta}(\frac{1}{2}(i+d), \frac{1}{2}(i+d))\right\}^i}. \tag{A.13}$$

In the same way as in Theorem 1, it can be shown that:

$$\begin{aligned} \log(E(|\boldsymbol{R}|^{\frac{1}{k}})) &= (k-1)\log 2 + \frac{2}{k}\sum_{i=1}^{k-1} i \Bigg\{ \log\left(\frac{1}{2}(i+d) + s_i\right) - \frac{1}{i+d+2s_i} \\ &\qquad - \log(i+d+2s_i) + \frac{1}{2(i+d+2s_i)} + O\left(\frac{1}{i^2}\right) \Bigg\} \\ &= -\frac{k-1}{k} + \frac{1}{k}\sum_{i=1}^{k-1} O\left(\frac{1}{i}\right), \end{aligned} \tag{A.14}$$

where $\exists s_i \in (0, 1/k)$ by the mean value theorem. Thus, it follows that:

$$\lim_{k\to\infty} \log(E(|\boldsymbol{R}|^{\frac{1}{k}})) = -1, \tag{A.15}$$

and also that:

$$\lim_{k\to\infty} E(|\boldsymbol{R}|^{\frac{1}{k}}) = \exp(-1). \tag{A.16}$$

In the same way, it can be shown that $\lim_{k\to\infty} E(|\boldsymbol{R}|^{\frac{2}{k}}) = \exp(-2)$. Thus,

$$\lim_{k\to\infty} \mathrm{Var}(|\boldsymbol{R}|^{\frac{1}{k}}) = \exp(-2) - (\exp(-1))^2 = 0. \tag{A.17}$$

These results imply that the effective dependence for a Wishart distribution converges to $1 - \exp(-1) = 0.632$ in probability. $\square$