Wishart Distributions and Inverse-Wishart Sampling

Stanley Sawyer — Washington University — Vs. April 30, 2007

1. Introduction. The Wishart distribution $W(\Sigma, d, n)$ is a probability distribution of random nonnegative-definite $d \times d$ matrices that is used to model random covariance matrices. The parameter n is the number of degrees of freedom, and Σ is a nonnegative-definite symmetric $d \times d$ matrix that is called the *scale matrix*. By definition

$$W \;\approx\; W(\Sigma, d, n) \;\approx\; \sum_{i=1}^{n} X_i X_i', \qquad X_i \;\approx\; N(0, \Sigma) \tag{1.1}$$

so that $W \approx W(\Sigma, d, n)$ is the distribution of a sum of n rank-one matrices defined by independent normal $X_i \in R^d$ with $E(X) = 0$ and $\mathrm{Cov}(X) = \Sigma$. In particular

$$E(W) \;=\; nE(X_i X_i') \;=\; n\,\mathrm{Cov}(X_i) \;=\; n\Sigma$$

(See `multivar.tex` for more details.) In general, any $X \approx N(\mu, \Sigma)$ can be represented

$$\begin{aligned} X = \mu + AZ, \qquad Z \;\approx\; N(0, I_d), \quad \text{so that} \\ \Sigma = \mathrm{Cov}(X) = A\,\mathrm{Cov}(Z)A' = AA' \end{aligned} \tag{1.2}$$

The easiest way to find A in terms of Σ is the LU-decomposition, which finds a unique lower diagonal matrix A with $A_{ii} \geq 0$ such that $AA' = \Sigma$. Then by (1.1) and (1.2) with $\mu = 0$

$$\begin{aligned} W(\Sigma, d, n) \;&\approx\; \sum_{i=1}^{n} (AZ_i)(AZ_i)' \;\approx\; A\left(\sum_{i=1}^{n} Z_i Z_i'\right)A', \quad Z_i \approx N(0, I_d) \\ &\approx\; A\,W(d, n)\,A' \qquad \text{where } W(d, n) = W(I_d, d, n) \end{aligned} \tag{1.3}$$

In particular, $W(\Sigma, d, n)$ can be easily represented in terms of $W(d, n) = W(I_d, d, n)$.

Assume in the following that $n > d$ and Σ is invertible. Then the density of the random $d \times d$ matrix W in (1.1) can be written

$$f(w, n, \Sigma) = \frac{|w|^{(n-d-1)/2} \exp\bigl(-(1/2)\,\mathrm{tr}(w\Sigma^{-1})\bigr)}{2^{dn/2}\pi^{d(d-1)/4}|\Sigma|^{n/2} \prod_{i=1}^{d} \Gamma\bigl((n+1-i)/2\bigr)} \tag{1.4}$$

where $|w| = \det(w)$, $|\Sigma| = \det(\Sigma)$, and $f(w, n, \Sigma) = 0$ unless w is symmetric and positive definite (Anderson 2003, Section 7.2, page 252).

2. The Inverse-Wishart Conjugate Prior. An important use of the Wishart distribution is as a conjugate prior for multivariate normal sampling. This leads to a d-dimensional analog of the inverse-gamma-normal conjugate prior for normal sampling in one dimension.

The likelihood function of n independent observations $X_i \approx N(\mu, \Sigma)$ for a $d \times d$ positive definite matrix Σ is

$$\begin{aligned} L(\mu, \Sigma, X) &= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\bigl(-(1/2)(X_i - \mu)'\Sigma^{-1}(X_i - \mu)\bigr) \\ &= \frac{1}{(2\pi)^{nd/2}|\Sigma|^{n/2}} \exp\left(-(1/2)\sum_{i=1}^{n}(X_i - \mu)'\Sigma^{-1}(X_i - \mu)\right) \qquad (2.1) \end{aligned}$$

The sum in (2.1) can be written

$$\begin{aligned} &\sum_{i=1}^{n} \sum_{a=1}^{d}\sum_{b=1}^{d}(X_{ia} - \mu_a)(\Sigma^{-1})_{ab}(X_{ib} - \mu_b) \\ &= \sum_{a=1}^{d}\sum_{b=1}^{d}(\Sigma^{-1})_{ab}\sum_{i=1}^{n}(X_{ia} - \mu_a)(X_{ib} - \mu_b) \\ &= \sum_{a=1}^{d}\sum_{b=1}^{d}(\Sigma^{-1})_{ab}Q(\mu)_{ab} = \operatorname{tr}\bigl(\Sigma^{-1}Q(\mu)\bigr) \qquad (2.2) \end{aligned}$$

where

$$\begin{aligned} Q(\mu) &= \sum_{i=1}^{n}(X_i - \mu)(X_i - \mu)' \\ &= \sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})' \;+\; n(\overline{X} - \mu)(\overline{X} - \mu)' \\ &= Q_0 + n\nu\nu', \qquad \nu = \overline{X} - \mu \qquad (2.3) \end{aligned}$$

Substituting (2.3) into (2.2) and (2.1) leads to the expression

$$\begin{aligned} L(\mu, \Sigma, X) &= \frac{\exp\bigl(-(1/2)\operatorname{tr}\bigl(Q_0\Sigma^{-1}\bigr)\bigr)\exp\bigl(-(1/2)n\nu'\Sigma^{-1}\nu\bigr)}{(2\pi)^{nd/2}|\Sigma|^{n/2}} \\ &= C_{nd}\,|\Sigma^{-1}|^{(n-1)/2}\exp\bigl(-(1/2)\operatorname{tr}\bigl(Q_0\Sigma^{-1}\bigr)\bigr) \\ &\quad \times \frac{1}{\sqrt{2\pi|\Sigma|}}\exp\bigl(-(1/2)n(\mu - \overline{X})'\Sigma^{-1}(\mu - \overline{X})\bigr) \qquad (2.4) \end{aligned}$$

where

$$Q_0 \;=\; \sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})' \tag{2.5}$$

Note that the integral $\int L(\mu,\Sigma,X)d\mu$ in (2.4) is the same as the Wishart density (1.4) with Σ^{-1} replaced by w, Q_0 in (2.4) replaced by Σ^{-1} in (1.4) so that $Q_0\Sigma^{-1}$ is replaced by $w\Sigma^{-1}$, and n replaced by $n-d$, within multiplicative constants that depend only on n, d, and X.

The similarity in forms between (1.4) and the first factor in (2.4) suggests that we might be able to sample from the density $L(\mu,\Sigma,X)$ in (2.4) by generating random variables by

$$\begin{aligned}
&\text{(i)} \quad W \approx W(d,n,Q_0^{-1}) \\
&\text{(ii)} \quad \Sigma = W^{-1} \\
&\text{(iii)} \quad \mu = \overline{X} + (A/\sqrt{n})Z, \quad \Sigma = AA', \quad Z \approx N(0,I_d)
\end{aligned} \tag{2.6}$$

One subtlety is that the density of Σ in (2.6) will not be $f(n,Q_0^{-1},\Sigma^{-1})$ for $f(n,S,W)$ in (1.4), or at least will not be this density with respect to Lebesgue measure $d\Sigma$ in R^{d^2}. In general, for any function $\phi(y)\ge 0$,

$$\begin{aligned}
E\big(\phi(\Sigma)\big) \;&=\; E\big(\phi(W^{-1})\big) \;=\; \int \phi(y^{-1})f(n,Q_0^{-1},y)\,dy \\
&=\; \int \phi(y)f(n,Q_0^{-1},y^{-1})\,d(y^{-1}) \\
&=\; \int \phi(y)f(n,Q_0^{-1},y^{-1})J_y(y^{-1})\,dy
\end{aligned}$$

where $J_y(y^{-1})$ is the absolute value of the Jacobian matrix of $y\to y^{-1}$.

Among all invertible matrices, a space of dimension d^2, the Jacobian $J_y(y^{-1}) = |y|^{-2d}$ (see Theorem 4.1 below). However, $f(w,n,\Sigma)$ in (1.4) is derived using the "Bartlett decomposition" (see Section 3 below) to parametrize positive definite symmetric matrices by a flat space of dimension $d(d+1)/2$. Anderson (2003) states $J_y(y^{-1}) = |y|^{-d-1}$ for the mapping $y\to y^{-1}$ restricted to symmetric matrices, but refers only to a theorem in his Appendix (Theorem A.4.6) that has only the d^2-dimensional result.

In any event, substituting $J_y(y^{-1}) = |y|^{-d-1}$ above leads to

$$E\big(\phi(\Sigma)\big) \;=\; \int \phi(y)|y|^{-d-1}f(n,Q_0^{-1},y^{-1})\,dy$$

Thus by (1.4) the joint density of (μ,Σ) generated by (2.6) is

$$\begin{aligned}
g(\mu,\Sigma) \;=\; &C|\Sigma|^{-(n+d+1)/2}\exp\big(-(1/2)\operatorname{tr}(\Sigma^{-1}Q_0)\big) \\
&\times \frac{1}{\sqrt{2\pi|\Sigma|}}\exp\big(-(1/2)n(\mu-\overline{X})'\Sigma^{-1}(\mu-\overline{X})\big)
\end{aligned} \tag{2.7}$$

The first factor in (2.7) is called the inverse-Wishart distribution by Anderson (2003, Theorem 7.7.1).

Gelman *et al.* (2003) define the four-parameter *inverse-Wishart-normal density* for (μ, Σ) as the density generated by

$$\begin{aligned} \Sigma^{-1} &\approx W(\nu_0, d, \Lambda_0^{-1}) \\ \mu \mid \Sigma &\approx N(\mu_0, \Sigma/\kappa_0) \end{aligned} \tag{2.8}$$

where κ_0, ν_0 are positive numbers, μ_0 is a real number, and Λ_0 is a $d \times d$ positive definite matrix. They also state the density

$$\begin{aligned} p(\mu, \Sigma) = |\Sigma|^{-((\nu_0+d+1)/2)} \exp\bigl(-(1/2)\,\mathrm{tr}(\Lambda_0 \Sigma^{-1})\bigr) \\ \times |\Sigma|^{-1/2} \exp\bigl(-(\kappa_0/2)(\mu - \mu_0)'\Sigma^{-1}(\mu - \mu_0)\bigr) \end{aligned} \tag{2.9}$$

for (2.8), which is the same as (2.7) with $\nu_0 = n$, $\Lambda_0 = Q_0$, $\kappa_0 = n$, and $\mu_0 = \overline{X}$.

Gelman *et al.* (2003) say that updating (2.8) or (2.9) with respect to an independent multivariate normal sample $X_1, X_2, \ldots, X_n$ with distribution $N(\mu, \Sigma)$ preserves the distribution with μ_0 etc. replaced by

$$\begin{aligned} \mu_n &= \frac{n}{\kappa_0 + n}\overline{X} + \frac{\kappa_0 \mu_0}{\kappa_0 + n} \\ \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \Lambda_n &= \Lambda_0 + Q_0 + \frac{\kappa_0 n}{\kappa_0 + n}(\overline{X} - \mu_0)(\overline{X} - \mu_0)' \end{aligned} \tag{2.10}$$

for Q_0 in (2.5). Since ν_0 appears in (2.10) only in the additive update $\nu_n = \nu_0 + n$, the initial power of Σ in the density does not matter.

Gelman *et al.* also discuss using a Jeffrey's prior

$$\pi_0(\mu, \Sigma) = C/|\Sigma|^{(d+1)/2}$$

This would amount to increasing n by $d + 1$ in (2.6) or (2.7) or ν_0 by $d + 1$ in (2.9). This would at least guarantee that the random matrices W in (2.6) were always invertible.

3. A Fast Way to Generate Wishart-Distributed Random Variables. Suppose that we want to estimate parameters in a model with independent multivariate normal variables. Bayesian methods based on Gibbs sampling using (2.4)–(2.4) and (2.6) or (2.8) depend on being able to simulate Wishart-distributed random matrices in an efficient manner.

If we simulate $W \approx W(\Sigma, d, n)$ using the basic definition (1.1)–(1.3), then we have to generate nd independent standard normal random variables and use of order nd^2 operations for each simulated value of W. Odell and Feiveson (1966) (referenced in Liu, 2001) developed a way to simulate W in $O(d^2)$ operations, which is a considerable improvement in time if $n \gg d$.

Theorem 3.1 (Odell and Feiveson, 1966). Suppose that V_i $(1 \le i \le d)$ are independent random variables where V_i has a chi-square distribution with $n-i+1$ degrees of freedom (so that $n-d+1 \le n-i+1 \le n$). Suppose that N_{ij} are independent normal random variables with mean zero and variance one for $1 \le i < j \le d$, also independent of the V_i. Define random variables b_{ij} for $1 \le i, j \le d$ by $b_{ji} = b_{ij}$ for $1 \le i < j \le d$ and

$$\begin{aligned} b_{ii} &= V_i + \sum_{r=1}^{i-1} N_{ri}^2, \qquad 1 \le i \le d \\ b_{ij} &= N_{ij}\sqrt{V_i} + \sum_{r=1}^{i-1} N_{ri}N_{rj}, \qquad i < j \le d \end{aligned} \tag{3.1}$$

Then $B = \{b_{ij}\}$ has a Wishart distribution $W(d, n) = W(I_d, d, n)$.

Remarks. (1) We assume that empty sums in (3.1) are zero, so that (3.1) implies $b_{11} = V_1$ and $b_{1j} = N_{1j}\sqrt{V_1}$. Note that each diagonal entry b_{ii} is individually chi-square with n degrees of freedom.

Odell and Feiveson state the theorem with $n-1$ in place of n, so that V_i is chi-squared with $n-i$ degrees of freedom (from $n-d$ to $n-1$) and the conclusion is that $B \approx W(n-1, d)$. They suggest that their algorithm was used for simulation studies of regression filters for estimating spacecraft trajectories.

The random variables b_{ij} in (3.1) can be defined in a single double loop. If V_i is defined in an outer loop just before b_{ii} is defined, and N_{ij} in an inner loop just before b_{ij} is defined, then, by induction on i, the values N_{ri}, N_{rj} in (3.1) have been previously defined. In particular, B in (3.1) can be defined in a single double loop without requiring a preliminary double loop to define V_i and N_{ij}, although storage would have to be allocated for previous values of N_{ij}.

(2) Define a random matrix T by

$$\begin{aligned} &T_{ij} = N_{ji} \quad (1 \le j < i \le d) \\ &T_{ii} = \sqrt{V_i}\,, \qquad T_{ij} = 0 \quad (i < j \le d) \end{aligned} \tag{3.2}$$

Then T is upper diagonal and (3.1) is equivalent to

$$b_{ij} = \sum_{r=1}^{\min(i,j)} T_{ir}T_{jr} \qquad \text{or}$$
$$B = TT' \qquad (B = \{b_{ij}\}) \tag{3.3}$$

Note that $T_{ir} \neq 0$ only for $r \leq i$ means that the only nonzero terms are on the diagonal or *above* the diagonal, since $(i, r) = (1, 1)$ is the upper left-hand corner of a matrix as usually written.

In general, the relation $B = TT'$ gives a one-one mapping between positive definite matrices B and upper diagonal matrices T with positive elements on the diagonal. Anderson (2003, Chapter 7) uses this fact to derive the formula (1.4) above for the Wishart density with a $d(d+1)/2$-dimensional parametrization of symmetric matrices. The proof of Theorem 3.1 in Anderson is essentially the same as the following (which follows Odell and Feiveson) except that Anderson has slightly more cryptic notation. Anderson calls $B = TT'$ the *Bartlett decomposition* after Bartlett (1939), although Anderson also attributes the term *rectangular coordinates* for T to Mahalanobis, Bose, and Roy (1937).

The Bartlett decomposition (3.2)-(3.3) also implies the following result, which can be used to give the exact distribution of the sample generalized variance.

Corollary 3.1. If $W \approx W(\Sigma, d, n)$ as in (1.1), then the random determinant

$$|W| \;\approx\; |\Sigma| \prod_{i=0}^{d-1} V_i, \tag{3.4}$$

where V_i are independent chi-square with $V_i \approx \chi^2_{n-i}$.

Proof of Corollary 3.1. $W \approx W(\Sigma, d, n) \approx ABA'$ by (1.3) above where A is deterministic, $AA' = \Sigma$, and $B \approx W(d, n)$. Thus $|W| = |ABA'| = |B||AA'| = |\Sigma||B|$. By (3.3), $|B| = |TT'| = |T|^2 = \prod_{i=1}^{d} t_{ii}^2 = \prod_{i=1}^{d} V_i$, so that (3.4) follows from Theorem 3.1

Proof of Theorem 3.1. As a first step, we represent the matrix entries of the random variable $B \approx W(d, n)$ as

$$b_{ab} = \sum_{i=1}^{n} Z_{ia}Z_{ib}, \qquad Z_{ia} \text{ independent } N(0, 1)$$

This can be written as the inner product $b_{ab} = (Z_a, Z_b)$ if the column vectors $Z_a = \{Z_{ia}\}$ are viewed as random vectors in R^n. The Gram-Schmidt orthogonalization of the vectors $\{Z_a\}$ is

$$Y_a \;=\; Z_a \;-\; \sum_{b=1}^{a-1} Y_b\,(Z_a, Y_b)/(Y_b, Y_b), \quad 1 \le a \le d \tag{3.5}$$

or in terms of the individual components

$$Y_{ia} \;=\; Z_{ia} \;-\; \sum_{b=1}^{a-1} Y_{ib}\left(\sum_{j=1}^{n} Z_{jb}Y_{jb}\right) \Big/ \sum_{j=1}^{n} Y_{jb}^2$$

By induction on a, $(Y_r, Y_a) = 0$ for $r < a \le d$. Expanding (Y_a, Y_a) in (3.5) into four terms,

$$\begin{aligned}
(Y_a, Y_a) \;&=\; (Z_a, Z_a) \;-\; \sum_{b=1}^{a-1} (Z_a, Y_b)^2/(Y_b, Y_b) && (3.6)\\
&=\; \sum_{i=1}^{n}\sum_{j=1}^{n} Z_{ia}\left(\delta_{ij} - \sum_{b=1}^{a-1} Y_{ib}Y_{jb}/(Y_b, Y_b)\right) Z_{ja} \\
&=\; \sum_{i=1}^{n}\sum_{j=1}^{n} Z_{ia}(R_a)_{ij}Z_{ja} && (3.7)
\end{aligned}$$

where R_a and Q_b are the $n \times n$ random matrices

$$(R_a)_{ij} \;=\; \delta_{ij} \;-\; \sum_{b=1}^{a-1} (Q_b)_{ij}, \qquad (Q_b)_{ij} = Y_{ib}Y_{jb}/(Y_b, Y_b) \tag{3.8}$$

Since $\{Y_a\}$ are orthogonal for $1 \le a \le d$, the Q_b in (3.8) are $n \times n$ rank-one random matrices with

$$Q_b = Q_b' = Q_b^2, \qquad Q_bQ_c = 0, \quad 1 \le c < b \le d$$

Thus Q_b are random orthogonal projection matrices, also orthogonal to one another. For the same reason, R_a in (3.8) is a random orthogonal projection matrix with rank $n - (a-1) = n + 1 - a$ in R_n.

The matrices R_a in (3.7)–(3.8) are random, but by induction depend on Z_{jb} only for $1 \le b < a \le d$. Conditional on $\{Z_{jb}\}$ for $b < a$, R_a is a deterministic orthogonal projection matrix of rank $n + 1 - a$. Thus the quadratic form $(Y_a, Y_a) = (Z_a, R_aZ_a)$ in (3.7) has a chi-squared distribution with $n + 1 - a$ degrees of freedom. Since this distribution is the same for all values of $\{Z_{jb}\}$ for $b < a$, it follows that the absolute distribution of

$V_a = (Z_a, R_a Z_a)$ is chi-square with $n+1-a$ degrees of freedom and that it is independent of Z_{jb} for $b < a$.

It follows from the same argument that $(Z_a, Q_b Z_a) = (Z_a, Y_b)^2/(Y_b, Y_b)$ for $b < a$ are independent chi-square random variables with one degree of freedom, which are also independent of $V_a = (Z_a, R_a Z_a)$. Thus we can define independent standard normal random variables

$$N_{ba} = \pm\sqrt{(Z_a, Q_b Z_a)} = \pm(Z_a, Y_b)/\sqrt{(Y_b, Y_b)}, \qquad 1 \le b < a \le d \tag{3.9}$$

which are also independent of V_a. Similarly, the families $\{V_a, N_{ba}\}$ are independent for different a since $\{V_a, N_{ba}\}$ have a fixed distribution conditional on $\{Z_{jc}\}$ for $c < a$. It only remains to relate the coefficients

$$b_{ab} \;=\; \sum_{i=1}^{n} Z_{ia} Z_{ib} \;=\; (Z_a, Z_b) \;\approx\; W(d, n)$$

to V_a and N_{bc}. First, by (3.6) and (3.9)

$$b_{aa} \;=\; (Z_a, Z_a) \;=\; (Y_a, Y_a) + \sum_{b=1}^{a-1} (Z_a, Y_b)^2/(Y_b, Y_b) \;=\; V_a \;+\; \sum_{b=1}^{a-1} N_{ba}^2$$

This proves the first part of (3.1). By (3.5)

$$\begin{aligned} b_{ba} &= (Z_a, Z_b) \\ &= \left(Y_a + \sum_{c=1}^{a-1} Y_c (Z_a, Y_c)/(Y_c, Y_c),\;\; Y_b + \sum_{e=1}^{b-1} Y_e (Z_b, Y_e)/(Y_e, Y_e) \right) \end{aligned}$$

Since Y_a on the left side of the large inner product is orthogonal to both terms on the right side for $b < a$,

$$b_{ba} \;=\; (Z_a, Y_b) + \sum_{e=1}^{b-1} (Z_a, Y_e)(Z_b, Y_e)/(Y_e, Y_e) \;=\; N_{ba}\sqrt{V_b} \;+\; \sum_{c=1}^{b-1} N_{ca} N_{cb}$$

since $(Z_a, Y_b) = \pm N_{ba}\sqrt{(Y_b, Y_b)} = \pm N_{ba}\sqrt{V_b}$ for $b < a$ by (3.9). This completes the proof of (3.1).

4. The Jacobian of the Inverse of a Matrix. The purpose of this section is to prove

Theorem 4.1. Let $A = \{a_{ik}\}$ be an invertible $d \times d$ matrix, which we can view as a vector $A \in R^{d^2}$ by the encoding

$$A_I = a_{ik}, \quad I = id + k \quad \text{where} \quad 0 \le i, k < d, \quad 0 \le I < d^2$$

Then the mapping $A \to A^{-1}$ has the Jacobian matrix

$$\frac{\partial}{\partial A}(A^{-1}) \;=\; -(A')^{-1} \otimes A^{-1} \tag{4.1}$$

where $\otimes$ means a tensor product (see below). Moreover, if $|A|$ denotes the determinant in either R^d or R^{d^2},

$$\left|\frac{\partial}{\partial A}(A^{-1})\right| \;=\; -|A|^{-2d} \tag{4.2}$$

Remarks. **(1)** The Jacobian matrix on the left-hand side of (4.1) is $d^2 \times d^2$, while A, A', and A^{-1} are $d \times d$.

(2) In general, if $A = \{a_{ij}\}$ is $d_A \times d_A$ and $B = \{b_{k\ell}\}$ is a $d_B \times d_B$ matrix, the tensor product $A \otimes B$ is the $d_A d_B \times d_A d_B$ matrix with entries

$$(A \otimes B)_{IJ} = a_{ij} b_{k\ell}, \quad I = id_B + k, \;\; J = jd_B + \ell \tag{4.3}$$

In particular, if A is 2×2 and B is 5×5, then $A \otimes B$ is 10×10. The encoding (4.3) of pairs (i, k) into I (where i, j are “slow” indices and k, ℓ are “fast” indices) is equivalent to the block matrix form

$$A \otimes B \;=\; \begin{pmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{pmatrix}$$

Equation (4.2) in Theorem 4.1 follows from the identity

$$|A \otimes B| \;=\; |A|^{d_B} |B|^{d_A} \tag{4.4}$$

which is proven in Theorem 4.2 below.

(3) The identity (4.4) (Theorem 4.2) is Theorem A.4.5 in the Appendix of Anderson (2003). Theorem 4.1 is Theorem A.4.6. An analog of Theorem 4.1 for symmetric matrices ($J_y(y^{-1}) = |y|^{-d-1}$, since they inhabit a space of lower dimension; Anderson 2003, p272) is used to derive the inverse Wishart distribution and referred to Theorem A.4.6, which does not cover that case.

Proof of Theorem 4.1. Since $A^{-1}A = I_d$,

$$\sum_{b=1}^{d}(A^{-1})_{jb}A_{bc} = \delta_{jc}, \qquad 1 \le j, c \le d$$

Then

$$\begin{aligned}\sum_{b=1}^{d}\left(\frac{\partial}{\partial A_{ik}}(A^{-1})_{jb}\right)A_{bc} \;&=\; -\sum_{b=1}^{d}(A^{-1})_{jb}\frac{\partial}{\partial A_{ik}}A_{bc} \\ &=\; -\sum_{b=1}^{d}(A^{-1})_{jb}\delta_{ib}\delta_{kc} \;=\; -(A^{-1})_{ji}\delta_{kc} \end{aligned} \tag{4.5}$$

If we postmultiply both sides of (4.5) by $(A^{-1})_{c\ell}$ and sum over c, we obtain

$$\frac{\partial}{\partial A_{ik}}(A^{-1})_{j\ell} \;=\; -(A^{-1})_{ji}(A^{-1})_{k\ell}$$

The encoding (4.3) then implies

$$\frac{\partial}{\partial A_I}(A^{-1})_J \;=\; -\big((A')^{-1} \otimes A^{-1}\big)_{IJ}$$

which implies (4.1).

Theorem 4.2. Let $A = \{a_{ij}\}$ be a $d_A \times d_A$ positive-definite matrix and $B = \{b_{k\ell}\}$ a $d_B \times d_B$ positive definite matrix. Define $A \otimes B$, I, and J as in (4.3). Then

$$|A \otimes B| \;=\; |A|^{d_B}|B|^{d_A} \tag{4.6}$$

Proof. The proof uses the fact that a positive definite matrix can be written as $A = LU$ where L is upper diagonal and U is lower diagonal. We begin with four lemmas, some of whose proofs we leave as exercises.

Lemma 1. If $A = \{a_{ij}\}$ is upper or lower diagonal, then $|A|$ is the product of its diagonal elements. That is, $|A| = \prod_{i=1}^{d} a_{ii}$.

Lemma 2. If A and B are two upper diagonal matrices, then AB is also upper diagonal.

Lemma 3. If A and B are two upper diagonal matrices, then $A \otimes B$ is upper diagonal.

Proof. While Lemma 2 assumes that A and B are of the same dimension, Lemma 3 does not. Lemma 3 depends on how the indices $I = (i, k)$ are encoded, which we assume is as in (4.3). Thus $(A \otimes B)_{IJ} = a_{ik} b_{k\ell}$ for $I = id_B + k$ and $J = jd_B + \ell$. We need to show that $(A \otimes B)_{IJ} = 0$ unless $J \le I$. If $J > I$, then either $j > i$ or $j = i$ and $\ell > k$. If the first case $a_{ij} = 0$ and $b_{k\ell} = 0$ in the second case.

Lemma 4. If A and B and $d_A \times d_A$ matrices and C and D are $d_C \times d_C$, then

$$(A \otimes C)(B \otimes D) \;=\; AB \otimes CD \tag{4.7}$$

Proof. If $I = id_C + k$, $J = jd_C + \ell$, and $M = md_C + n$ as before,

$$\begin{aligned}
\big((A \otimes C)(B \otimes D)\big)_{IJ} \;&=\; \sum_{M=0}^{d_A d_C - 1} (A \otimes C)_{IM} (B \otimes D)_{MJ} \\
&=\; \sum_{m=0}^{d_A-1} \sum_{n=0}^{d_C-1} a_{im} c_{kn} b_{mj} d_{n\ell} \;=\; \sum_{m=0}^{d_A-1} a_{im} b_{mj} \sum_{n=0}^{d_C-1} c_{kn} d_{n\ell} \\
&=\; (AB)_{ij} (CD)_{k\ell} \;=\; (AB \otimes CD)_{IJ}
\end{aligned}$$

Proof of Theorem 4.2. Write $A = L_A U_A$ and $B = L_B U_B$ where L_A, L_B are upper diagonal and U_A, U_B are lower diagonal. Then by Lemma 4

$$A \otimes B \;=\; L_A U_A \otimes L_B U_B \;=\; (L_A \otimes L_B)(U_A \otimes U_B)$$

and

$$|A \otimes B| \;=\; |(L_A \otimes L_B)(U_A \otimes U_B)| \;=\; |L_A \otimes L_B|\,|U_A \otimes U_B| \tag{4.8}$$

Since $L_A \otimes L_B$ is upper diagonal by Lemma 3, then by Lemma 1

$$|L_A \otimes L_B| \;=\; \prod_{I=0}^{d_A d_B - 1} (L_A \otimes L_B)_{II} \;=\; \prod_{i=0}^{d_A-1} \prod_{j=0}^{d_B-1} (c_{ii} d_{jj})$$

if $L_A = \{c_{ik}\}$ and $L_B = \{d_{j\ell}\}$. Then

$$\begin{aligned}
|L_A \otimes L_B| \;&=\; \prod_{i=0}^{d_A-1} \left(c_{ii}^{d_B} \prod_{j=0}^{d_B-1} d_{jj} \right) \;=\; \left(\prod_{i=0}^{d_A-1} c_{ii} \right)^{d_B} \left(\prod_{j=0}^{d_B-1} d_{jj} \right)^{d_A} \\
&=\; |L_A|^{d_B} |L_B|^{d_A}
\end{aligned}$$

By the same argument $|U_A \otimes U_B| = |U_A|^{d_B} |U_B|^{d_A}$, and $|A| = |L_A U_A| = |L_A||U_A|$ and $|B| = |L_B U_B| = |L_B||U_B|$. Putting this together with (4.8) implies (4.6). This completes the proof of the theorem.

References.

1. Anderson, T. W. (2003) An introduction to multivariate statistical analysis, 3rd edn. John Wiley and Sons, New York.

2. Bartlett, M. S. (1939) A note on tests of significance in multivariate analysis. *Proceedings of the Cambridge Philosophical Society* **35,** 180–185.

3. Gelman, A, J. Carlin, H. Stern, and D. Rubin (2003) Bayesian data analysis, 2nd edition. Chapman & Hall/CRC, Boca Raton.

4. Liu, Jun S. (2001) Monte Carlo strategies in scientific computing. Springer Series in Statistics, Springer-Verlag.

5. Mahalanobis, P.C., R.C. Bose, and S.N. Roy (1937) Normalisation of statistical variates and the use of rectangular coordinates in the theory of sampling distributions. *Sankhya* **3,** 1–40.

6. Odell, P.L., and A.H. Feiveson (1966) A numerical procedure to generate a sample covariance matrix. *Jour. Amer. Stat. Assoc.* **61,** 199–203.