Lecture Notes on Multivariate Statistical Analysis

1 Outline

- Basic properties of multivariate normal distribution
- Jacobian computation and exterior products
- Wishart distribution : density, characteristic function
- Some properties of Wishart distribution
- Joint distribution of the eigenvalues of Wishart
- Maximum likelihood estimates of eigenvalues and eigenvectors

2 Multivariate normal distribution

Normal distribution is central to the classical multivariate statistical analysis. Here are a few important facts about multivariate normal random variables.

Definition 1: *A random variable X taking values in $\mathbb{R}^p$ is said to have (real) p-variate Normal distribution with mean μ and covariance Σ (a $p \times p$ positive semi-definite matrix), expressed as $X \sim N_p(\mu, \Sigma)$, if the characteristic function of X is given by*

$$\phi_X(\mathbf{t}) := \mathbb{E}\exp(i\mathbf{t}^T X) = \exp(i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}), \qquad \text{for all } \mathbf{t} \in \mathbb{R}^p. \tag{1}$$

If Σ is positive definite then the $N_p(\mu, \Sigma)$ distribution has a density function given by

$$f_X(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}(\det\Sigma)^{1/2}}\exp(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)), \qquad \mathbf{x} \in \mathbb{R}^p. \tag{2}$$

Below we list some useful properties of $N_p(\mu, \Sigma)$ distribution.

Proposition 1: *If X has $N_p(\mu, \Sigma)$ distribution, then $AX + b$ has $N_m(A\mu + b, A\Sigma A^T)$ distribution for any $m \times p$ matrix A and any $m \times 1$ vector b.*

Proposition 2: *If X is $N_p(\mu, \Sigma)$ (with $p \geq 2$) and X, μ, and Σ are partitioned as*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where X_1, μ_1 are $k \times 1$ and Σ_{11} is $k \times k$ with $1 \leq k < p$, then X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.

Proof : Show that the characteristic function of X factors into a product of characteristic functions of $N_k(\mu_1, \Sigma_{11})$ and $N_{p-k}(\mu_2, \Sigma_{22})$ random variables (and hence these are independent) if and only if $\Sigma_{12} = 0$.

Proposition 3: *Let X be $N_p(\mu, \Sigma)$ (with $p \geq 2$) and X, μ, and Σ are partitioned as*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where X_1, μ_1 are $k \times 1$ and Σ_{11} is $k \times k$ with $1 \leq k < p$. Let Σ_{22}^- be a generalized inverse of Σ_{22}, i.e. a matrix satisfying

$$\Sigma_{22}\Sigma_{22}^-\Sigma_{22} = \Sigma_{22}, \tag{3}$$

and let $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-\Sigma_{21}$. Then

(a) $X_1 - \Sigma_{12}\Sigma_{22}^-X_2$ has $N_k(\mu_1 - \Sigma_{12}\Sigma_{22}^-\mu_2, \Sigma_{11\cdot 2})$ distribution and is independent of X_2, and

(b) the conditional distribution of X_1 given X_2 is $N_k(\mu_1 + \Sigma_{12}\Sigma_{22}^-(X_2 - \mu_2), \Sigma_{11\cdot 2})$.

Proof : First, note that since Σ is non-negative definite, $\mathcal{R}(\Sigma_{12}) \subset \mathcal{R}(\Sigma_{22})$ where $\mathcal{R}$ denotes the "row space", i.e. the linear span of the rows. Therefore, $\Sigma_{12} = B\Sigma_{22}$ for some $k \times (p-k)$ matrix B. This implies that

$$\Sigma_{12}\Sigma_{22}^-\Sigma_{22} = B\Sigma_{22}\Sigma_{22}^-\Sigma_{22} = B\Sigma_{22} = \Sigma_{12}$$

by (3). Let

$$C = \begin{bmatrix} I_k & -\Sigma_{12}\Sigma_{22}^- \\ 0 & I_{p-k} \end{bmatrix}.$$

Then CX is of the form

$$CX = \begin{bmatrix} X_1 - \Sigma_{12}\Sigma_{22}^-X_2 \\ X_2 \end{bmatrix}$$

and has a p-variate normal distribution with mean

$$\begin{bmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^-\mu_2 \\ \mu_2 \end{bmatrix}$$

and covariance matrix (check)

$$C\Sigma C^T = \begin{bmatrix} \Sigma_{11\cdot 2} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}.$$

Then (a) follows from *Proposition 1* and *Proposition 2*, and from this and *Proposition 1*, (b) follows by conditioning on X_2.

Proposition 4: *If X has $N_p(\mu, I_p)$ distribution then $Z = X^T X = \| X \|^2$ has the χ^2 distribution with p degrees of freedom and noncentrality parameter $\delta = \| \mu \|^2$ (written $Z \sim \chi_p^2(\delta)$), with density*

$$f_{p,\delta}(z) = e^{-\delta/2} \ {}_0F_1(\frac{p}{2}; \frac{\delta z}{4}) \frac{1}{2^{p/2}\Gamma(p/2)} z^{n/2-1} e^{-z/2}, \qquad z > 0, \tag{4}$$

where

$${}_0F_1(a; w) = \sum_{k=0}^{\infty} \frac{w^k}{(a)_k k!}$$

where $(a)_k = a(a+1)\cdots(a+k-1)$ for $k \geq 1$ (and $(a)_0 = 1$), and $a, w \in \mathbb{C}$.

Corollary 1: *If X is $N_p(\mu, \Sigma)$ and Σ is non-singular, then for any $b \in \mathbb{R}^p$, the random variable $(X-b)^T \Sigma^{-1} (X-b)$ has $\chi_p^2(\delta)$ distribution, with $\delta = (\mu - b)^T \Sigma^{-1} (\mu - b)$.*

Proposition 5: *The characteristic function of $\chi_p^2(\delta)$ random variable is given by*

$$\phi_{p,\delta}(t) = \exp\left(\frac{i\delta t}{1-2it}\right)(1-2it)^{-p/2}, \qquad t \in \mathbb{R}. \tag{5}$$

Proposition 6: *If X is $N_p(\mu, I_p)$ and B is a $p \times p$ symmetric matrix then $X^T B X$ has a noncentral χ^2 distribution if and only if B is idempotent (i.e., $B^2 = B$), in which case the degrees of freedom and the noncentrality parameters are respectively $k = rank(B) = \ trace(B)$ and $\delta = \mu^T B \mu$.*

Proof : Suppose that B is idempotent of rank k. Then $B = H\Lambda H^T$ for some $p \times p$ orthogonal matrix H, where

$$\Lambda = \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}.$$

Therefore, $X^T B X = X^T H \Lambda H^T X = \sum_{j=1}^{k} V_j^2$ where $V = H^T X$ has $N_p(H^T\mu, I_p)$ distribution (since $H^T H = I_p$). Then V_j's are independent $N(\nu_j, 1)$ random variables with $\nu_j = h_j^T \mu$ (where h_j is the j-th column of H). The "if part" follows now by *Proposition 4* after noticing that $\sum_{j=1}^{k} \nu_j^2 = \sum_{j=1}^{k} \mu^T h_j h_j^T \mu = \mu^T B \mu = \delta$.

To prove the "only if" part, start with the spectral decomposition $B = H\Lambda H^T$ where Λ is diagonal with the diagonal elements $\lambda_1 \geq \ldots \geq \lambda_p$. Then by arguments as above, it follows that $X^T B X = \sum_{j=1}^{p} \lambda_j V_j^2$ where V_j are independent $N(\nu_j, 1)$ for $j = 1, \ldots, p$. Now computing characteristic function of $\sum_{j=1}^{p} \lambda_j V_j^2$ by means of *Proposition 5* and comparing with the characteristic function of a $\chi_k^2(\delta)$ random variable for $\delta = \mu^T B \mu = \sum_{j=1}^{p} \nu_j^2$, it follows that we must have rank$(B) = k$ and $\lambda_j = 1$ for $j = 1, \ldots, k$ and $\lambda_j = 0$ for $j > k$, which concludes the proof.

3 Jacobians and exterior products

One of the key ingredients in the study of multivariate distributions is the computation of Jacobians for appropriate transformations. Let $X \in \mathbb{R}^m$ be a random vector with probability density function $f(\mathbf{x})$ with support $S \subset \mathbb{R}^m$. Suppose that the transformation $\mathbf{y} = \mathbf{y}(\mathbf{x}) = (y_1(\mathbf{x}), \ldots, y_m(\mathbf{x}))$ is 1-1 on S onto $T \subset \mathbb{R}^m$. Assuming that the partial derivatives $\partial x_i/\partial y_j$ $(i, j = 1, \ldots, m)$ exist and are continuous on T (viewing $\mathbf{x} = \mathbf{x}(\mathbf{y})$ as the inverse function of $\mathbf{y}$ on T onto S), the density function of $Y = \mathbf{y}(X)$ is given by

$$g(\mathbf{y}) = f(\mathbf{x}(\mathbf{y}))|J(\mathbf{x} \to \mathbf{y})|, \qquad \mathbf{y} \in T, \tag{6}$$

where $J(\mathbf{x} \to \mathbf{y})$ is the Jacobian of the transformation from $\mathbf{x}$ to $\mathbf{y}$, i.e.,

$$J(\mathbf{x} \to \mathbf{y}) = \det \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \frac{\partial x_m}{\partial y_1} & \cdots & \frac{\partial x_m}{\partial y_m} \end{bmatrix} = \det\left(\frac{\partial x_i}{\partial y_j}\right). \tag{7}$$

In multivariate analysis it is often useful to do the Jacobian computation in terms of (coordinate-wise) differentials of smooth transform. Hence we introduce, following James (1954) the idea of *wedge products* of differentials. A detailed account can be found in Muirhead (1980).

Definition 2: *An exterior differential form of degree $r \geq 1$ in $\mathbb{R}^m$ is an expression of the type*

$$\sum_{i_1 \leq \cdots \leq i_r} h_{i_1,\ldots,i_r}(\mathbf{x}) dx_{i_1} \wedge \cdots \wedge dx_{i_r}, \tag{8}$$

where $i_1, \ldots, i_r$ are indices from the set $\{1, \ldots, m\}$, $h_{i_1,\ldots,i_r}(\mathbf{x})$ are analytic functions of $\mathbf{x} = (x_1, \ldots, x_m)$, and the "wedge product" $dx_{i_1} \wedge \cdots \wedge dx_{i_r}$ is skew-symmetric, i.e. (for $r \geq 2$) if $1 \leq k < l \leq r$, then

$$\begin{aligned} & dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}} \wedge dx_{i_k} \wedge dx_{i_{k+1}} \wedge \cdots \wedge dx_{i_{l-1}} \wedge dx_{i_l} \wedge dx_{i_{l+1}} \wedge \cdots \wedge dx_{i_r} \\ = \; & -dx_{i_1} \wedge \cdots \wedge dx_{i_{k-1}} \wedge dx_{i_l} \wedge dx_{i_{k+1}} \wedge \cdots \wedge dx_{i_{l-1}} \wedge dx_{i_k} \wedge dx_{i_{l+1}} \wedge \cdots \wedge dx_{i_r}. \end{aligned} \tag{9}$$

3.1 Implications

Let $r = 1$, and consider the 1-1 continuously differentiable transformation $\mathbf{x} \to \mathbf{y}$. Then the ordinary differential

$$dx_i = \frac{\partial x_i}{\partial y_1} dy_1 + \cdots + \frac{\partial x_i}{\partial y_m} dy_m, \qquad (1 \leq i \leq m) \tag{10}$$

is an exterior differential form of degree 1 in $\mathbb{R}^m$.

The skew-symmetry condition (9) implies in particular that if in the wedge product $dx_{i_1} \wedge \cdots \wedge dx_{i_r}$ any term is repeated then the expression must be zero. Consequently we have the following:

(a) A form of degree m has only one term, namely $h(\mathbf{x})dx_1 \wedge \cdots \wedge dx_m$.

(b) A form of degree greater than m is zero.

(c) For $1 \leq r \leq m$ the expression (8) reduces to

$$\sum_{i_1 < \cdots < i_r} h_{i_1,\ldots,i_r}(\mathbf{x}) dx_{i_1} \wedge \cdots \wedge dx_{i_r}. \tag{11}$$

We can regard (11) as an integrand in an r-dimensional surface integral in $\mathbb{R}^m$. To illustrate this point, and to understand the definition of wedge products, we consider the simple case when $m = 2$. The aim is to show that using the definition we can recover the well known change-of-variable formula

$$\int_A f(x_1, x_2) dx_1 dx_2 = I = \int_{A'} f(x_1(\mathbf{y}), x_2(\mathbf{y})) \det\left(\frac{\partial x_i}{\partial y_j}\right) dy_1 dy_2, \tag{12}$$

where A' is the image of A under the (smooth) 1-1 transformation $\mathbf{x} \to \mathbf{y}$.

Formally, we can multiply the expressions for dx_1 and dx_2 given (10) to obtain

$$I = \int_{A'} f(\mathbf{x}(\mathbf{y})) \left(\frac{\partial x_1}{\partial y_1} dy_1 + \frac{\partial x_1}{\partial y_2} dy_2\right)\left(\frac{\partial x_2}{\partial y_1} dy_1 + \frac{\partial x_2}{\partial y_2} dy_2\right). \tag{13}$$

We verify that if we treat the products of differentials appearing in (13) as a wedge product then we recover (12). In order that this holds, we must have

$$\left(\frac{\partial x_1}{\partial y_1} dy_1 + \frac{\partial x_1}{\partial y_2} dy_2\right) \wedge \left(\frac{\partial x_2}{\partial y_1} dy_1 + \frac{\partial x_2}{\partial y_2} dy_2\right) = \left(\frac{\partial x_1}{\partial y_1}\frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2}\frac{\partial x_2}{\partial y_1}\right) dy_1 \wedge dy_2. \tag{14}$$

Simplifying the left hand side by using the associative and distributive laws and applying the rule $dy_1 \wedge dy_2 = -dy_2 \wedge dy_1$ and $dy_1 \wedge dy_1 = dy_2 \wedge dy_2 = 0$ (following (9)) we have the identity (14).

3.2 Some important transformations

In this section we compute the Jacobians for some important transformations using the idea of exterior products (i.e. products of exterior differential forms). We shall use $\wedge_{i=1}^m dx_i$ to denote the exterior product $dx_1 \wedge \cdots \wedge dx_m$. A key result is the following:

Proposition 7: *If* $d\mathbf{y}$ *is an* $m \times 1$ *vector of differentials and if* $d\mathbf{x} = Bd\mathbf{y}$ *where* B *is an* $m \times m$ *nonsingular matrix (so that* $\mathbf{x}$ *is a vector of linear differential forms), then*

$$\wedge_{i=1}^m dx_i = (\det B) \wedge_{i=1}^m dy_i. \tag{15}$$

Proof : By using associative and distributive rules the left side can be written as

$$\wedge_{i=1}^m dx_i = p(B) \wedge_{i=1}^m dy_i,$$

for some polynomial $p(B)$ in the elements of the matrix B. The following properties hold.

(i) $p(B)$ is linear in each row of B.

(ii) By the skew-symmetry condition (9) if two factors dx_i and dx_j are interchanged then the sign of $\wedge_{i=1}^m dx_i$ is reversed. But interchanging the position of dx_i and dx_j corresponds to interchanging the i-th and j-th rows of B. So interchanging any two rows of B reverses the sign of $p(B)$.

(iii) $p(I_m) = 1$.

But properties (i), (ii) and (iii) characterize the determinant of a matrix. Hence $p(B) = \det B$.

We now introduce some notations to compute Jacobians for transformations of matrices. So, for an $n \times m$ matrix X, let dX denote the matrix of differentials $((dx_{ij}))_{1 \le i \le m, 1 \le j \le n}$. It follows that if X is $n \times m$ and Y is $m \times p$ then

$$d(XY) = X \cdot dY + dX \cdot Y.$$

For an arbitrary $n \times m$ matrix X, we shall use (dX) to denote the exterior product of the mn elements of dX:

$$(dX) = \wedge_{j=1}^m \wedge_{i=1}^n dx_{ij}.$$

If X is a *symmetric* $m \times m$ matrix, the symbol (dX) will denote the exterior product of the $\frac{1}{2}m(m+1)$ *distinct* elements of dX:

$$(dX) = \wedge_{1 \le i \le j \le m} dx_{ij}.$$

If X is $m \times m$ lower-triangular, then

$$(dX) = \wedge_{1 \le j \le i \le m} dx_{ij}.$$

If X is $m \times m$ skew symmetric (i.e. $X^T = -X$), then as a convention, define (dX) in terms of the $\frac{1}{2}m(m-1)$ distinct elements below the diagonal

$$(dX) = \wedge_{1 \le j < i \le m} dx_{ij}.$$

Proposition 8: *If $X = BY$ where X and Y are $n \times m$ matrices and B is a fixed nonsingular $n \times n$ matrix, then*

$$(dX) = (\det B)^m (dY),$$

so that $J(X \to Y) = (\det B)^m$.

Proposition 9: *If $X = BYC$ where X and Y are $n \times m$ matrices and B and C are nonsingular matrices of dimension $n \times n$ and $m \times m$, respectively, then*

$$(dX) = (\det B)^m (\det C)^n (dY),$$

so that $J(X \to Y) = (\det B)^m (\det C)^n$.

Proposition 10: *If* $X = BYB^T$*, where* X *and* Y *are* $m \times m$ *symmetric matrices and* B *is an* $m \times m$ *nonsingular matrix, then*

$$(dX) = (\det B)^{m+1}(dY).$$

If instead Y *is skew-symmetric (i.e.,* $Y^T = -Y$*), then*

$$(dX) = (\det B)^{m-1}(dY).$$

Proof : Since $X = BYB^T$, it implies that $dX = BdYB^T$, from which easily that

$$(dX) = p(B)(dY) \tag{16}$$

for some polynomial $p(B)$ in the elements of B. Now verify using (16), with B replaced by B_1B_2 that the polynomial satisfies the equation

$$p(B_1B_2) = p(B_1)p(B_2) \tag{17}$$

for arbitrary $m \times m$ matrices B_1 and B_2. The only polynomials in the elements of a square matrix satisfying (17) for all B_1 and B_2 are ineteger powers of $\det B$ (see MacDuffee (1943), Chapter 3), so that

$$p(B) = (\det B)^k \qquad \text{for some integer } k.$$

Now to calculate k choose a special form of B. Choose $B = \mathrm{diag}(b, 1, \ldots, 1)$, so that

$$BYB^T = \begin{bmatrix} b^2y_{11} & by_{12} & \cdots & by_{1m} \\ by_{21} & y_{22} & \cdots & y_{2m} \\ \cdot & \cdot & \cdots & \cdot \\ by_{m1} & y_{m2} & \cdots & y_{mm} \end{bmatrix}.$$

Taking exterior products of elements on and above diagonal for the case when Y is symmetric (and of elements only above the diagonal when Y is skew symmetric), we see that

$$(BdYB^T) = \begin{cases} b^{m+1}(dY) & \text{if } \ Y = Y^T \\ b^{m-1}(dY) & \text{if } \ Y = -Y^T. \end{cases}$$

Proposition 11: *If* A *is an* $m \times m$ *positive definite matrix and* $A = TT^T$ *where* T *is a lower triangular matrix with positive diagonal elements, then*

$$(dA) = 2^m \prod_{i=1}^{m} t_{ii}^{m+1-i}(dT).$$

Proposition 12: *If $X = Y^{-1}$, where Y is a $m \times m$ symmetric matrix, then*

$$(dX) = (\det Y)^{-(m+1)}(dY).$$

The following proposition gives the Jacobian for the Gram-Schmidt orthogonalization (or QR decomposition) of an $n \times m$ matrix.

Proposition 13 : *Let Z be an $n \times m$ matrix ($n \geq m$) of rank m. Let $Z = H_1 T$ where H_1 is an $n \times m$ matrix with $H_1^T H_1 = I_m$ and T is an $m \times m$ upper triangular matrix with positive diagonal elements. Let H_2 (a function of H_1) be an $n \times (n-m)$ matrix such that $H = [H_1 : H_2]$ is an $n \times n$ orthogonal matrix. Write $H = [\mathbf{h}_1 : \cdots : \mathbf{h}_m : \mathbf{h}_{m+1} : \cdots : \mathbf{h}_n]$ where $\mathbf{h}_1, \ldots, \mathbf{h}_m$ are columns of H_1 and $\mathbf{h}_{m+1}, \ldots, \mathbf{h}_n$ are columns of H_2. Then*

$$(dZ) = \prod_{i=1}^{m} t_{ii}^{n-i}\ (dT)\ (H_1^T dH_1),$$

where

$$(H_1^T dH_1) = \wedge_{i=1}^{m} \wedge_{j=i+1}^{n} \mathbf{h}_j^T d\mathbf{h}_i.$$

Moreover, if $A = Z^T Z$, then

$$(dZ) = 2^{-m} (\det A)^{(n-m-1)/2}\ (dA)\ (H_1^T dH_1).$$

Proposition 14: *For the following transformation from rectangular coordinates $x_1, \ldots, x_m$ to polar coordinates $r, \theta_1, \ldots, \theta_{m-1}$ in $\mathbb{R}^m$:*

$$\begin{aligned} x_1 &= r \sin\theta_1 \cdots \sin\theta_{m-2} \sin\theta_{m-1} \\ x_j &= r \prod_{k=1}^{m-j} \sin\theta_k \cos\theta_{m-j+1}, \quad j = 2, \ldots, m \\ &\qquad (r > 0, \quad 0 < \theta_j \leq \pi \text{ for } 1 \leq j \leq m-2, \quad 0 < \theta_{m-1} \leq 2\pi), \end{aligned} \tag{18}$$

we have

$$\wedge_{j=1}^{m} dx_j = r^{m-1} \prod_{j=1}^{m-2} (\sin\theta_j)^{m-j-1} (\wedge_{j=1}^{m-1} d\theta_j) \wedge dr$$