

A TUTORIAL ON
MULTIVARIATE
STATISTICAL ANALYSIS

Craig A. Tracy
UC Davis

SAMSI
September 2006

ELEMENTARY STATISTICS

Collection of (real-valued) data from a sequence of experiments

$$X_1, X_2, \dots, X_n$$

Might make assumption underlying law is $N(\mu, \sigma^2)$ with unknown mean μ and variance σ^2 . Want to estimate μ and σ^2 from the data.

Sample Mean & Sample Variance:

$$\bar{X} = \frac{1}{n} \sum_j X_j, \quad S = \frac{1}{n-1} \sum_j (X_j - \bar{X})^2$$

Estimators are “unbiased”

$$\mathbb{E}(\bar{X}) = \mu, \quad \mathbb{E}(S) = \sigma^2$$

Theorem: If X_1, X_2, \dots are independent $N(\mu, \sigma^2)$ variables then \bar{X} and S are independent. We have that \bar{X} is $N(\mu, \sigma^2/n)$ and $(n-1)S/\sigma^2$ is $\chi^2(n-1)$.

Recall $\chi^2(d)$ denotes the chi-squared distribution with d degrees of freedom. Its density is

$$f_{\chi^2}(x) = \frac{1}{2^{d/2} \Gamma(d/2)} x^{d/2-1} e^{-x/2}, \quad x \geq 0,$$

where

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt, \quad \Re(z) > 0.$$

MULTIVARIATE GENERALIZATIONS

From the classic textbook of Anderson[1]:

Multivariate statistical analysis is concerned with data that consists of sets of measurements on a number of individuals or objects. The sample data may be heights and weights of some individuals drawn randomly from a population of school children in a given city, or the statistical treatment may be made on a collection of measurements, such as lengths and widths of petals and lengths and widths of sepals of iris plants taken from two species, or one may study the scores on batteries of mental tests administered to a number of students.

p = # of sets of measurements on a given individual,

n = # of observations = sample size

Remarks:

- In above examples, one can assume that $p \ll n$ since typically many measurements will be taken.
- Today it is common for $p \gg 1$, so n/p is no longer necessarily large.

Vehicle Sound Signature Recognition: Vehicle noise is a stochastic signal. The power spectrum is discretized to a vector of length $p = 1200$ with $n \approx 1200$ samples from the same kind of vehicle.

Astrophysics: Sloan Digital Sky Survey typically has many observations (say of quasar spectrum) with the spectra of each quasar binned resulting in a large p .

Financial data: S&P 500 stocks observed over monthly intervals for twenty years.

GAUSSIAN DATA MATRICES

The data are now n independent column vectors of length p

$$\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$$

from which we construct the $n \times p$ data matrix

$$X = \begin{pmatrix} \longleftarrow & \vec{x}_1^T & \longrightarrow \\ \longleftarrow & \vec{x}_2^T & \longrightarrow \\ & \vdots & \\ \longleftarrow & \vec{x}_n^T & \longrightarrow \end{pmatrix}$$

The Gaussian assumption is that

$$\vec{x}_j \sim N_p(\mu, \Sigma)$$

Many applications assume the mean has been already subtracted out of the data, i.e. $\mu = 0$.

Multivariate Gaussian Distribution

If x and y are vectors, the matrix $x \otimes y$ is defined by

$$(x \otimes y)_{jk} = x_j y_k$$

If $\mu = \mathbb{E}(x)$ is the mean of the random vector x , then the **covariance matrix** of x is the $p \times p$ matrix

$$\Sigma = \mathbb{E}[(x - \mu) \otimes (x - \mu)]$$

Σ is a symmetric, non-negative definite matrix. If $\Sigma > 0$ (positive definite) and $X \sim N_p(\mu, \Sigma)$, then the density function of X is

$$f_X(x) = (2\pi)^{-p/2} (\det \Sigma)^{-1/2} \exp \left[-\frac{1}{2} (x - \mu, \Sigma^{-1}(x - \mu)) \right], \quad x \in \mathbb{R}^p$$

Sample mean:

$$\bar{x} = \frac{1}{n} \sum_j \vec{x}_j, \quad \mathbb{E}(\bar{x}) = \mu$$

Sample covariance matrix:

$$S = \frac{1}{n-1} \sum_{j=1}^n (\vec{x}_j - \bar{x}) \otimes (\vec{x}_j - \bar{x})$$

For $\mu = 0$ the sample covariance matrix can be written simply as

$$\frac{1}{n-1} X^T X$$

Some Notation: If X is a $n \times p$ data matrix formed from the n independent column vectors x_j , $\text{cov}(x_j) = \Sigma$, we can form one column vector $\text{vec}(X)$ of length pn

$$\text{vec}(X) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

The covariance of $\text{vec}(X)$ is the $np \times np$ matrix

$$I_n \otimes \Sigma = \begin{pmatrix} \Sigma & 0 & 0 & \cdots & 0 \\ 0 & \Sigma & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \Sigma \end{pmatrix}$$

In this case we say the data matrix X constructed from n independent $x_j \sim N_p(\mu, \Sigma)$ has distribution

$$N_p(M, I_n \otimes \Sigma)$$

where $M = \mathbb{E}(X) = \mathbf{1} \otimes \mu$, $\mathbf{1}$ is the column vector of all 1's.

WISHART DISTRIBUTION

Definition: If $A = X^T X$ where the $n \times p$ matrix X is $N_p(0, I_n \otimes \Sigma)$, $\Sigma > 0$, then A is said to have *Wishart distribution* with n degrees of freedom and covariance matrix Σ . We will say A is $W_p(n, \Sigma)$.

Remarks:

- The Wishart distribution is the multivariate generalization of the chi-squared distribution.
- $A \sim W_p(n, \Sigma)$ is positive definite with probability one if and only if $n \geq p$.
- The sample covariance matrix,

$$S = \frac{1}{n-1} A$$

is $W_p(n-1, \frac{1}{n-1} \Sigma)$.

WISHART DENSITY FUNCTION, $n \geq p$

Let \mathcal{S}_p denote the space of $p \times p$ positive definite (symmetric) matrices. If $A = (a_{jk}) \in \mathcal{S}_p$, let

$$(dA) = \text{volume element of } A = \bigwedge_{j \leq k} da_{jk}$$

The *multivariate gamma function* is

$$\Gamma_p(a) = \int_{\mathcal{S}_p} e^{-\text{tr}(A)} (\det A)^{a-(p+1)/2} (dA), \Re(a) > (p-1)/2.$$

Theorem: If A is $W_p(n, \Sigma)$ with $n \geq p$, then the density function of A is

$$\frac{1}{2^{np} \Gamma_p(n/2) (\det \Sigma)^{n/2}} e^{-\frac{1}{2} \text{tr}(\Sigma^{-1} A)} (\det A)^{(n-p-1)/2}$$

Sketch of Proof:

- The density function for X is the multivariate Gaussian (including volume element (dX))

$$(2\pi)^{-np/2} (\det \Sigma)^{-n/2} e^{-\frac{1}{2}\text{tr}(\Sigma^{-1} X^T X)} (dX)$$

- Recall the *QR factorization* [8]: Let X denote an $n \times p$ matrix with $n \geq p$ with full column rank. Then there exists a unique $n \times p$ matrix Q , $Q^T Q = I_p$, and a unique $n \times p$ upper triangular matrix R with positive diagonal elements so that $X = QR$. Note $A = X^T X = R^T R$.

- A Jacobian calculation [1, 13]: If $A = X^T X$, then

$$(dX) = 2^{-p} (\det A)^{(n-p-1)/2} (dA)(Q^T dQ)$$

where

$$(Q^T dQ) = \bigwedge_{j=1}^p \bigwedge_{k=j+1}^n q_k^T dq_j$$

and $Q = (q_1, \dots, q_p)$ is the column representation of Q .

- Thus the joint distribution of A and Q is

$$(2\pi)^{-np/2} (\det \Sigma)^{-n/2} e^{-\frac{1}{2} \text{tr}(\Sigma^{-1} A)} \times \\ 2^{-p} (\det A)^{(n-p-1)/2} (dA)(Q^T dQ)$$

- Now integrate over all Q . Use fact that

$$\int_{\mathcal{V}_{n,p}} (Q^T dQ) = \frac{2^p \pi^{np/2}}{\Gamma_p(n/2)}$$

and $\mathcal{V}_{n,p}$ is the set of real $n \times p$ matrices Q satisfying $Q^T Q = I_p$. (When $n = p$ this is the orthogonal group.)

Remarks regarding the Wishart density function

- Case $p = 2$ obtain by R. A. Fisher in 1915.
- General p by J. Wishart in 1928 by geometrical arguments.
- Proof outlined above came later. (See [1, 13] for complete proof.)
- When Q is a $p \times p$ orthogonal matrix

$$\frac{\Gamma_p(p/2)}{2^p \pi^{p^2/2}} (Q^T dQ)$$

is *normalized Haar measure* for the orthogonal group $\mathcal{O}(p)$. We denote this Haar measure by (dQ) .

- Siegel proved (see, e.g. [13])

$$\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma\left(a - \frac{1}{2}(j-1)\right)$$

EIGENVALUES OF A WISHART MATRIX

Theorem: If A is $W_p(n, \Sigma)$ with $n \geq p$ the joint density function for the eigenvalues ℓ_1, \dots, ℓ_p of A is

$$\frac{\pi^{p^2/2} 2^{-np/2} (\det \Sigma)^{-n/2}}{\Gamma_p(p/2) \Gamma_p(n/2)} \prod_{j=1}^p \ell_j^{(n-p-1)/2} \prod_{j < k} |\ell_j - \ell_k| \times$$

$$\int_{\mathcal{O}(p)} e^{-\frac{1}{2} \text{tr}(\Sigma^{-1} Q L Q^T)} (dQ), \quad (\lambda_1 > \dots > \lambda_p)$$

where $L = \text{diag}(\ell_1, \dots, \ell_p)$ and (dQ) is normalized Haar measure. Note that $\Delta(\ell) := \prod_{j < k} (\ell_j - \ell_k)$ is the Vandermonde.

Corollary: If A is $W_p(n, I_p)$, then the integral over the orthogonal group in the previous theorem is

$$e^{-\frac{1}{2} \sum_j \ell_j}.$$

Proof: Recall that the Wishart density function (times the volume element) is

$$\frac{1}{2^{np} \Gamma_p(n/2) (\det \Sigma)^{n/2}} e^{-\frac{1}{2} \text{tr}(\Sigma^{-1} A)} (\det A)^{(n-p-1)/2} (dA)$$

The idea is to diagonalize A by an orthogonal transformation and then integrate over the orthogonal group thereby giving the density function for the eigenvalues of A .

Let $\ell_1 > \dots > \ell_p$ be the ordered eigenvalues of A .

$$A = QLQ^T, \quad L = \text{diag}(\ell_1, \dots, \ell_p), \quad Q \in \mathcal{O}(p)$$

The j^{th} column of Q is a normalized eigenvector of A . The transformation is not 1-1 since $Q = [\pm q_1, \dots, \pm q_p]$ works for each fixed A . The transformation is made 1-1 by requiring that the 1^{st} element of each q_j is nonnegative. This restricts Q (as A varies) to a 2^{-p} part of $\mathcal{O}(p)$. We compensate for this at the end.

We need an expression for the volume element (dA) in terms of Q and L . First we compute the differential of A

$$\begin{aligned}
 dA &= dQ L Q^T + Q dL Q^T + Q L dQ^T \\
 Q^T dA Q &= Q^T dQ L + dL + L dQ^T Q \\
 &= -dQ^t Q dL + L dQ^T Q + dL \\
 &= [L, dQ^T Q] + dL
 \end{aligned}$$

(We used $Q^T Q = I$ implies $Q^T dQ = -dQ^T Q$.)

We now use the following fact (see, e.g., page 58 in [13]): If $X = BYB^T$ where X and Y are $p \times p$ symmetric matrices, B is a nonsingular $p \times p$ matrix, then $(dX) = (\det B)^{p+1}(dY)$. In our case Q is orthogonal so the volume element (dA) equals the volume element ($Q^T dA Q$). The volume element is the exterior product of the diagonal elements of $Q^T dA Q$ times the exterior product of the elements above the diagonal.

Since L is diagonal, the commutator

$$[L, dQ^T Q]$$

has zero diagonal elements. Thus the exterior product of the diagonal elements of $Q^T dA Q$ is $\bigwedge_j d\ell_j$.

The exterior product of the elements coming from the commutator is

$$\prod_{j < k} (\ell_j - \ell_k) \bigwedge_{j < k} q_k^T dq_j$$

and so

$$\begin{aligned} (dA) &= \bigwedge_{j < k} q_k^T dq_j \Delta(\ell) \bigwedge_j d\ell_j \\ &= \frac{2^p \pi^{p^2/2}}{\Gamma_p(p/2)} (dQ) \Delta(\ell) \bigwedge_j d\ell_j \end{aligned}$$

The theorem now follows once integrate over all of $\mathcal{O}(p)$ and divide the result by 2^p .

- One is interested in *limit laws* as $n, p \rightarrow \infty$. For $\Sigma = I_p$, Johnstone [11] proved, using RMT methods, for centering and scaling constants

$$\begin{aligned}\mu_{np} &= (\sqrt{n-1} + \sqrt{p})^2, \\ \sigma_{np} &= (\sqrt{n-1} + \sqrt{p}) \left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}} \right)^{1/3}\end{aligned}$$

that

$$\frac{\ell_1 - \mu_{np}}{\sigma_{np}}$$

converges in distribution as $n, p \rightarrow \infty$, $n/p \rightarrow \gamma < \infty$, to the GOE largest eigenvalue distribution [15].

- El Karoui [6] has extended the result to $\gamma \leq \infty$. The case $p \gg n$ appears, for example, in microarray data.
- Soshnikov [14] has lifted Gaussian assumption under the additional restriction $n - p = O(p^{1/3})$.

- For $\Sigma \neq I_p$, the difficulty in establishing limit theorems comes from the integral

$$\int_{\mathcal{O}(p)} e^{-\frac{1}{2} \text{tr}(\Sigma^{-1} Q \Lambda Q^T)} (dQ)$$

Using zonal polynomials infinite series expansions have been derived for this integral, but these expansions are difficult to analyze. See Muirhead [13].

- For *complex* Gaussian data matrices X similar density formulas are known for the eigenvalues of $X^* X$. Limit theorems for $\Sigma \neq I_p$ are known since the analogous group integral, now over the unitary group, is known explicitly—the Harish Chandra–Itzykson–Zuber (HCIZ) integral (see, e.g. [17]). See the work of Baik, Ben Arous and P  ch   [2, 3] and El Karoui [7].

PRINCIPAL COMPONENT ANALYSIS (PCA),

H. Hotelling, 1933

Population Principal Components: Let x be a $p \times 1$ random vector with $\mathbb{E}(x) = \mu$ and $\text{cov}(x) = \Sigma > 0$. Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ denote the eigenvalues of Σ and H an orthogonal matrix diagonalizing Σ : $H^T \Sigma H = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$. We write H in column vector form

$$H = [h_1, \dots, h_p]$$

so that h_j is the $p \times 1$ eigenvector of Σ corresponding to eigenvalue λ_j . Define the $p \times 1$ vector

$$u = H^T x = (u_1, \dots, u_p)^T$$

$$\begin{aligned} \text{then } \text{cov}(u) &= \mathbb{E} \left((H^T x - H^T \mu) \otimes (H^T x - H^T \mu) \right) \\ &= H^T \mathbb{E} \left((x - \mu) \otimes (x - \mu) \right) H \\ &= H^T \Sigma H = \Lambda \Rightarrow u_j \text{ uncorrelated.} \end{aligned}$$

Definition: u_j is called the j^{th} *principal component* of x . Note $\text{var}(u_j) = \lambda_j$.

Statistical interpretations: The claim is that u_1 is that linear combination of components of x that has *maximum variance*.

Proof: For simplicity of notation, set $\mu = 0$. Let b denote any $p \times 1$ vector, $b^T b = 1$, and form $b^T x$.

$$\text{var}(b^T x) = \mathbb{E}(b^T x \cdot b^T x) = \mathbb{E}(b^T x \cdot (b^T x)^T) = b^T \mathbb{E}(x x^T) b = b^T \Sigma b.$$

We want to maximize the right hand side subject to the constraint $b^T b = 1$. By the method of Lagrange multipliers we maximize

$$b^T \Sigma b - \lambda(b^T b - 1)$$

Since Σ is symmetric the vector of partial derivatives is

$$2\Sigma b - 2\lambda b$$

Thus b must be an eigenvector with eigenvalue λ .

The largest variance corresponds to choosing the largest eigenvalue.

The general result is that u_r has maximum variance of all normalized combinations uncorrelated with u_1, \dots, u_{r-1} .

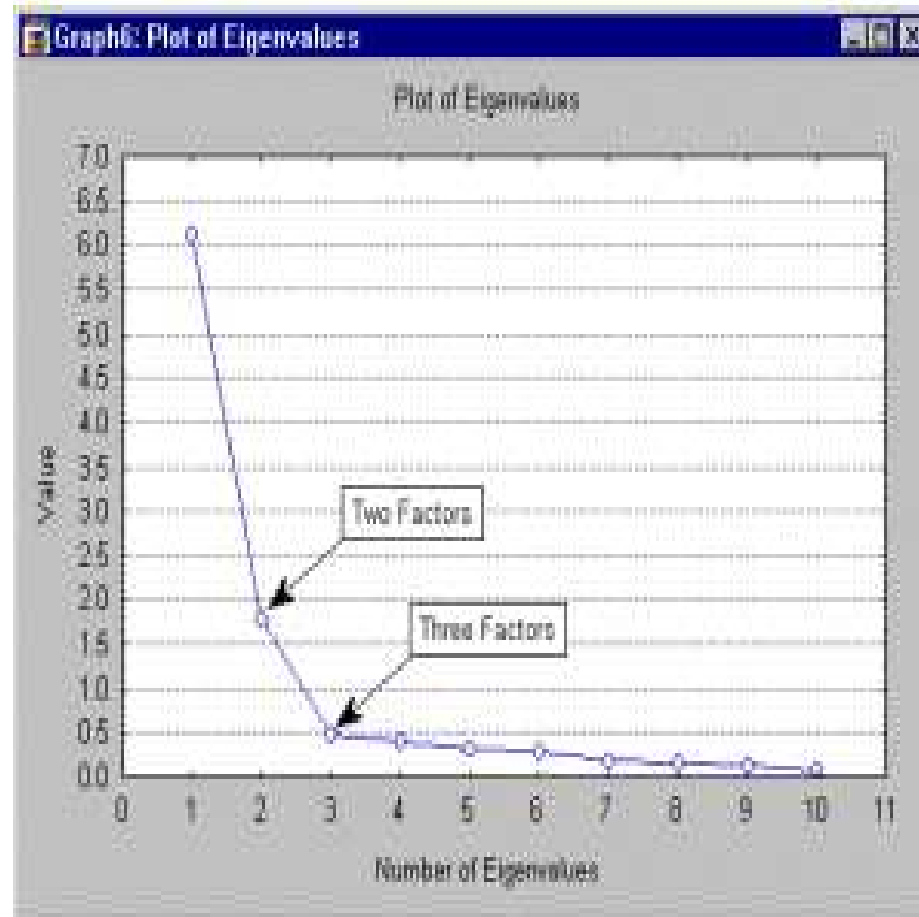
Sample principal components: Let S denote the sample covariance matrix of the data matrix X and let $Q = [q_1, \dots, q_p]$ a $p \times p$ orthogonal matrix diagonalizing S :

$$Q^T S Q = \text{diag}(\ell_1, \dots, \ell_p)$$

The ℓ_j are the *sample variances* that are estimates for λ_j . The vectors q_j are *sample estimates* for the vectors h_j .

If x is the random vector and $u = H^T x$ is the vector of principal components, then $\hat{u} = Q^T x$ is the vector of *sample principal components*.

SCREE PLOTS



In applications: How many of the ℓ_j 's are significant?

CANONICAL CORRELATION ANALYSIS (CCA)

H. Hotelling, 1936

Suppose a large data set is naturally decomposed into two groups. For example, $p \times 1$ random vectors $\vec{x}_1, \dots, \vec{x}_n$ make up one set and $q \times 1$ random vectors $\vec{y}_1, \dots, \vec{y}_m$ the other. We are interested in the correlations between these two data sets. For example, in medicine we might have n measurements of age, height, and weight ($p = 3$) and m measurements of systolic and diastolic blood pressures ($q = 2$). We are interested in what combination of the components of x is most correlated with a combination of the components of y .

Population Canonical Correlations: Let x and y be two random vectors of size $p \times 1$ and $q \times 1$, respectively. We assume $\mathbb{E}(x) = \mathbb{E}(y) = 0$ and $p \leq q$. Form the $(p + q) \times 1$ vector

$$\begin{pmatrix} x \\ y \end{pmatrix}$$

and its $(p + q) \times (p + q)$ covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Let

$$u := \alpha^T x \in \mathbb{R}, \quad v := \gamma^T y \in \mathbb{R}$$

where α and γ are vectors to be determined. We want to maximize the correlation

$$\text{corr}(u, v) = \frac{\text{cov}(u, v)}{\sqrt{\text{var}(u)\text{var}(v)}}$$

The correlation does not change under scale transformations $u \rightarrow cu$, etc. so we can maximize this correlation subject to the constraints

$$\mathbb{E}(u^2) = \mathbb{E}(\alpha^T x \cdot \alpha^T x) = \alpha^T \Sigma_{11} \alpha = 1 \quad (1)$$

$$\mathbb{E}(v^2) = \gamma^T \Sigma_{22} \gamma = 1 \quad (2)$$

Under these constraints

$$\text{corr}(u, v) = \mathbb{E}(\alpha^T x \cdot \gamma^T y) = \alpha^T \Sigma_{12} \gamma.$$

Let

$$\psi = \alpha^T \Sigma_{12} \gamma - \frac{1}{2} \rho (\alpha^T \Sigma_{11} \alpha - 1) - \frac{1}{2} \lambda (\gamma^T \Sigma_{22} \gamma - 1)$$

where ρ and λ are Lagrange multipliers. Set the vector of partial derivatives to zero:

$$\frac{\partial \psi}{\partial \alpha} = \Sigma_{12} \gamma - \rho \Sigma_{11} \alpha = 0 \quad (3)$$

$$\frac{\partial \psi}{\partial \gamma} = \Sigma_{12}^T \alpha - \lambda \Sigma_{22} \gamma = 0 \quad (4)$$

If we left multiply (3) by α^T and (4) by γ^T , use the normalization conditions (1) and (2) we conclude $\lambda = \rho$. Thus (3) and (4) become

$$\begin{pmatrix} -\rho \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\rho \Sigma_{22} \end{pmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} = 0 \quad (5)$$

with $\text{corr}(u, v) = \rho$

and $\rho \geq 0$ is a solution to

$$\det \begin{pmatrix} -\rho \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\rho \Sigma_{22} \end{pmatrix} = 0$$

This is a polynomial in ρ of degree $(p + q)$. Let ρ_1 denote the maximum root and α_1 and γ_1 corresponding solutions to (5).

Definition: u_1 and v_1 are called the *first canonical variables* and their correlation $\rho_1 = \text{corr}(u_1, v_1)$ is called the *first canonical correlation coefficient*.

More generally, the *rth pair of canonical variables* is the pair of linear combinations $u_r = (\alpha^{(r)})^T x$ and $v_r = (\gamma^{(r)})^T y$, each of unit variance and uncorrelated with the first $r - 1$ pairs of canonical variables and having maximum correlation. The correlation $\text{corr}(u_r, v_r)$ is the *rth canonical correlation coefficient*.

REDUCTION TO AN EIGENVALUE PROBLEM

Since

$$\begin{pmatrix} -\rho\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\rho\Sigma_{22} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -\rho I & \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ \Sigma_{21} & -\rho I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

Thus the determinantal equation becomes

$$\det(\rho^2 I - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}) = 0$$

The nonzero roots ρ_1, \dots, ρ_k are called the *population canonical correlation coefficients*. $k = \text{rank}(\Sigma_{12})$.

In applications Σ is not known. One uses the *sample covariance matrix* S to obtain *sample canonical correlation coefficients*.

AN EXAMPLE

The first application of Hotelling's canonical correlations is by F. Waugh [16] in 1942. He begins his paper with

Professor Hotelling's paper, "Relations between Two Sets of Variates," should be widely known and his method used by practical statisticians. Yet, few practical statisticians seem to know of the paper, and perhaps those few are inclined to regard it as a mathematical curiosity rather than an important and useful method for analyzing concrete problems. This may be due to ...

The Practical Problem: Relation of wheat characteristics to flour characteristics. Guiding principle

The grade of the raw material should give a good indication of the probable grade of the finished product.

The data: 138 samples of Canadian Hard Red Spring wheat and the flour made from each these samples.^a

wheat quality

flour quality

$x_1 =$ kernel texture

$y_1 =$ wheat per barrels of flour

$x_2 =$ test weight

$y_2 =$ ash in flour

$x_3 =$ damaged kernels

$y_3 =$ crude protein in flour

$x_4 =$ crude protein in wheat

$y_4 =$ gluten quality index

$x_5 =$ foreign material

$u_1 = \alpha_1^T x =$ index of wheat quality, $v_1 = \gamma_1^T y =$ index of flour quality

$$\text{corr}(u_1, v_1) = 0.909$$

^aIn this example $p > q$. The data are normalized to mean 0 and variance 1.

DISTRIBUTION OF SAMPLE CANONICAL CORRELATION COEFFICIENTS

Give a sample of size n observations on $\begin{pmatrix} x \\ y \end{pmatrix}$ drawn from $N_{p+q}(\mu, \Sigma)$ and A the (unnormalized) sample covariance matrix. Then W is $W_{p+q}(n, \Sigma)$. We have [13]

Theorem (Constantine, 1963): Let A have the $W_{p+q}(n, \Sigma)$ distribution where $p \leq q$, $n \geq p + q$ and Σ and A are partitioned as above. Then the joint probability density function of r_1^2, \dots, r_p^2 , the eigenvalues of $A_{11}^{-1} A_{12} A_{22}^{-1} A_{21}$ (let $\xi_j := r_j^2$) is

$$c_{p,q,n} \prod_{j=1}^p (1 - \rho_j^2)^{n/2} \prod_{j=1}^p \left[\xi_j^{(q-p-1)/2} (1 - \xi_j)^{(n-p-q-1)/2} \right] \cdot |\Delta(\xi)| \cdot \mathcal{F}(\xi)$$

where

- $c_{p,q,n}$ is a normalization constant.
- $\Delta(\xi) = \text{Vandermonde determinant} = \prod_{j < k}^p (\xi_j - \xi_k)$.
- $\mathcal{F}(\xi)$ is a two-matrix hypergeometric function which can be expressed as an infinite series involving zonal polynomials. See Theorem 11.3.2 and Definition 7.3.2 in Muirhead [13].

Null Distribution: For $\Sigma_{12} = 0$ (x and y are independent), the above joint density for the sample canonical correlation coefficients reduces to

$$c_{p,q,n} \prod_{j=1}^p \left[\xi_j^{(q-p-1)/2} (1 - \xi_j)^{(n-p-q-1)/2} \right] \cdot |\Delta(\xi)|$$

In this case the distribution of the largest sample canonical correlation coefficient r_1 can be used for testing the null hypothesis: $H : \Sigma_{12} = 0$. We reject H for large values of r_1 .

ZONAL POLYNOMIALS

Zonal polynomials naturally arise in multivariate analysis when considering group integrals such as

$$\int_{\mathcal{O}(m)} e^{-\text{tr}(XHYH^T)} (dH)$$

where X and Y are $m \times m$ symmetric, positive definite matrices and (dH) is normalized Haar measure. Zonal polynomials can be defined either through the representation theory of $GL(m, \mathbb{R})$ [12] or as eigenfunctions of certain Laplacians [5, 13].

Let \mathcal{S}_m denote the space of $m \times m$ symmetric, positive definite matrices. Zonal polynomials, $C_\lambda(X)$, $X \in \mathcal{S}_m$ are certain homogeneous polynomials in the eigenvalues of X that are indexed by partitions λ .

To give the precise definition we need some preliminary definitions.

Definition: A *partition* λ of n is a sequence $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_\ell)$ where the $\lambda_j \geq 0$ are weakly decreasing and $\sum_j \lambda_j = n$. We denote this by $\lambda \vdash n$.

For example, $\lambda = (5, 3, 3, 1)$ is a partition of 12. The number of nonzero parts of λ is called the *length* of λ , denoted $\ell(\lambda)$.

If λ and μ are two partitions of n , we say $\lambda < \mu$ (*lexicographic order*) if, for some index i , $\lambda_j = \mu_j$ for $j < i$ and $\lambda_j < \mu_j$. For example

$$(1, 1, 1, 1) < (2, 1, 1) < (2, 2) < (3, 1) < (4)$$

If $\lambda \vdash n$ with $\ell(\lambda) = m$, we define the monomial

$$x^\lambda = x_1^{\lambda_1} x_2^{\lambda_2} \cdots x_m^{\lambda_m}$$

and say x^μ is of *higher weight* than x^λ if $\mu > \lambda$.

Metrics and Laplacians

Let $X \in \mathcal{S}_m$ and $dX = (dx_{jk})$ the matrix of differentials of X . We define a metric on \mathcal{S}_m by

$$(ds)^2 = \text{tr} (X^{-1}dX \cdot X^{-1}dX)$$

A simple computation shows this metric is invariant under

$$X \longrightarrow LXL^T, \quad L \in GL(m, \mathbb{R}).$$

Let $n = m(m + 1)/2 = \#$ of independent elements of X . We denote by $\text{vec}(X) \in \mathbb{R}^n$ the column vector representation of X , e.g.

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{pmatrix}, \quad x := \text{vec}(X) = \begin{pmatrix} x_{11} \\ x_{12} \\ x_{22} \end{pmatrix}$$

The metric $(ds)^2$ is a quadratic differential in dx .

For example,

$$(ds)^2 = \begin{pmatrix} dx_{11} & dx_{12} & dx_{22} \end{pmatrix} \cdot G(x) \cdot \begin{pmatrix} dx_{11} \\ dx_{12} \\ dx_{22} \end{pmatrix}$$

where

$$G(x) = (g_{ij}) = \frac{1}{(x_{11}x_{22} - x_{12}^2)^2} \begin{pmatrix} x_{22}^2 & -2x_{12}x_{22} & x_{12}^2 \\ -2x_{12}x_{22} & 2x_{11}x_{22} + 2x_{12}^2 & -2x_{11}x_{12} \\ x_{12}^2 & -2x_{11}x_{12} & x_{11}^2 \end{pmatrix}$$

Labeling $x = \text{vec}(X)$ with a single index the differential is in the standard form

$$(ds)^2 = \sum_{i < j} dx_i g_{ij} dx_j$$

and the *Laplacian* associated to this metric is

$$\Delta_X = (\det G)^{-1/2} \sum_{j=1}^n \frac{\partial}{\partial x_j} \left[(\det G)^{1/2} \sum_{i=1}^n g^{ij} \frac{\partial}{\partial x_i} \right]$$

where

$$G^{-1} = (g^{ij})$$

More succinctly, if

$$\nabla_X = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix},$$

$$\Delta_X = (\det G)^{-1/2} \left(\nabla_X, (\det G)^{1/2} G^{-1} \nabla_X \right)$$

where (\cdot, \cdot) is the standard inner product on \mathbb{R}^n .

One can show that Δ_X is an *invariant differential operator*:

$$\Delta_{LXL^T} = \Delta_X, \quad L \in GL(m, \mathbb{R})$$

We now diagonalize X

$$X = HYH^T, \quad Y = \text{diag}(y_1, \dots, y_m), \quad H \in \mathcal{O}(m).$$

The Laplacian Δ_X is now expressed in terms of a *radial part* and an *angular part*. The radial part of Δ_X is the differential operator

$$\sum_{j=1}^m y_j^2 \frac{\partial^2}{\partial y_j^2} + \sum_{j=1}^m \sum_{\substack{k=1 \\ k \neq j}}^m \frac{y_j^2}{y_j - y_k} \frac{\partial}{\partial y_j} + \sum_j y_j \frac{\partial}{\partial y_j}$$

We now let Δ_X denote only the radial part.

We can now define zonal polynomials!

Definition: Let $X \in \mathcal{S}_m$ with eigenvalues x_1, \dots, x_m and let $\lambda = (\lambda_1, \dots, \lambda - m) \vdash k$ into not more than m parts. Then $C_\lambda(X)$ is the symmetric, homogeneous polynomial of degree k in x_j such that

1. The term of highest weight in $C_\lambda(X)$ is x^λ
2. $C_\lambda(X)$ is an eigenfunction of the Laplacian Δ_X .
3.
$$(\operatorname{tr}(X))^k = (x_1 + \dots + x_m)^k = \sum_{\substack{\lambda \vdash k \\ \ell(\lambda) \leq m}} C_\lambda(X).$$

Remarks:

- Must show there is an unique polynomial satisfying these requirements.
- Eigenvalue in (2) equals $\alpha_\lambda := \sum_j \lambda_j(\lambda_j - j) + k(m + 1)/2$.
- Program MOPS [5] computes zonal polynomials.

The following theorem is at the core of why zonal polynomials appear in multivariate statistical analysis.

Theorem: If $X, Y \in S_p$, then

$$\int_{\mathcal{O}(p)} C_\lambda(XHYH^T) (dH) = \frac{C_\lambda(X)C_\lambda(Y)}{C_\lambda(I_p)} \quad (6)$$

where (dH) is normalized Haar measure.

Proof: Let $f_\lambda(Y)$ denote the left-hand side of (6) and $Q \in \mathcal{O}(p)$.

$f_\lambda(QYQ^T) = f_\lambda(Y)$ (let $H \rightarrow HQ$ in integral and use invariance of the measure). Thus f_λ is a symmetric function of the eigenvalues of Y . Since C_λ is homogeneous of degree $|\lambda|$, so is f_λ . Now apply the

Laplacian to f_λ .

$$\begin{aligned}
\Delta_Y f_\lambda(Y) &= \int_{\mathcal{O}(p)} \Delta_Y C_\lambda(XHYH^T) (dH) \\
&= \int \Delta_Y C_\lambda(X^{1/2}HYH^T X^{1/2}) (dH) \\
&= \int \Delta_Y C_\lambda(LYL^T) (dH) \quad (L = X^{1/2}H) \\
&= \int \Delta_{LYL^T} C_\lambda(LYL^T) (dH) \quad \text{invariance of } \Delta_Y \\
&= \alpha_\lambda \int C_\lambda(LYL^T) (dH) \\
&= \alpha_\lambda f_\lambda(Y)
\end{aligned}$$

By definition of $C_\lambda(Y)$ we have $f_\lambda(Y) = d_\lambda C_\lambda(Y)$. Since $f_\lambda(I_p) = C_\lambda(X)$, we find d_λ and the theorem follows.

Using Zonal Polynomials to Evaluate Group Integrals

$$\begin{aligned}
 & \int_{\mathcal{O}(p)} e^{-\rho \operatorname{tr}(XHYH^T)} (dH) \\
 &= \sum_{k=0}^{\infty} \frac{\rho^k}{k!} \int_{\mathcal{O}(p)} (\operatorname{tr}(XHYH^T))^k (dH) \\
 &= \sum_{k=0}^{\infty} \frac{\rho^k}{k!} \sum_{\substack{\lambda \vdash k \\ \ell(\lambda) \leq m}} \int_{\mathcal{O}(p)} C_{\lambda}(XHYH^T) (dH) \\
 &= \sum_{k=0}^{\infty} \frac{\rho^k}{k!} \sum_{\substack{\lambda \vdash k \\ \ell(\lambda) \leq m}} \frac{C_{\lambda}(X)C_{\lambda}(Y)}{C_{\lambda}(I_p)}
 \end{aligned}$$

References

- [1] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, third edition, John Wiley & Sons, Inc., 2003.
- [2] J. Baik, G. Ben Arous, and S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices, *Ann. Probab.* **33** (2005), 1643–1697.
- [3] J. Baik, Painlevé formulas of the limiting distributions for nonnull complex sample covariance matrices, *Duke Math. J.* **133** (2006), 205–235.
- [4] M. Dieng and C. A. Tracy, Application of random matrix theory to multivariate statistics, arXiv: math.PR/0603543.
- [5] I. Dumitriu, A. Edelman and G. Shuman, MOPS: Multivariate Orthogonal Polynomials (symbolically), arXiv:math-ph/0409066.
- [6] N. El Karoui, On the largest eigenvalue of Wishart matrices with identity covariance when n , p and p/n tend to infinity, arXiv:

math.ST/0309355.

- [7] N. El Karoui, Tracy-Widom limit for the largest eigenvalue of a large class of complex Wishart matrices, arXiv: math.PR/0503109.
- [8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, second edition, Johns Hopkins University Press, 1989.
- [9] H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. of Educational Psychology* **24** (1933), 417–441, 498–520.
- [10] H. Hotelling, Relations between two sets of variates, *Biometrika* **28** (1936), 321–377.
- [11] I. M. Johnstone, On the distribution of the largest eigenvalue in principal component analysis, *Ann. Stat.* **29** (2001), 295–327.
- [12] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, 2nd ed., Oxford University Press, 1995.
- [13] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, John

Wiley & Sons Inc., 1982.

- [14] A. Soshnikov, A note on universality of the distribution of the largest eigenvalue in certain sample covariance matrices, *J. Statistical Physics* **108** (2002), 1033–1056.
- [15] C. A. Tracy and H. Widom, On orthogonal and symplectic matrix ensembles, *Commun. Math. Phys.* **177** (1996), 727–754.
- [16] F. V. Waugh, Regression between sets of variables, *Econometrica* **10** (1942), 290–310.
- [17] P. Zinn-Justin and J.-B. Zuber, On some integrals over the $U(N)$ unitary group and their large N limit, *J. Phys. A: Math. Gen.* **36** (2003), 3173–3193.