北京大學
PEKING UNIVERSITY

Notes for **Applied Multivariate Statistical Analysis**

lectured by Fang Yao* LaTeXed by T.-Y. Li†

2019 Fall

Abstract

The main goal of this course is to study multivariate statistical techniques, their conventional and modern ideas, innovative statistical methods, and novel computational tools, as well as new applications. The topics covered include Multivariate Modeling and Inferences, Model Assessment, Principal Components, Dimension Reduction, Classification, Cluster Analysis, if time permits, Nonparametric Methods, Support Vector Machines, Latent Variable Models, Artificial Neural Networks and so on.

 Please feel free to leave a message at https://zhuanlan.zhihu.com/p/91593024 (under construction). Any comments are welcome.

*Homepage: http://www.math.pku.edu.cn/teachers/yaof/
†E-Mail: kellty@pku.edu.cn

Contents

References

[JW] Richard A. Johnson, & Dean W. Wichern (2007). *Applied Multivariate Statistical Analysis* (6th ed.). Pearson Prentice Hall. http://www.mathstat.ualberta.ca/~wiens/stat575/datasets

[I] Alan J. Izenman (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer-Verlag. https://doi.org/10.1007/978-0-387-78189-1

[JWHT] Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani (2013). *An Introduction to Statistical Learning, with Applications in R*. Springer Science+Business Media. https://doi.org/10.1007/978-1-4614-7138-7

[G] Hui-Xuan Gao (2005). *Applied Multivariate Statistical Analysis* (in Chinese). Peking University Press.

[HS] Wolfgang K. Härdle, & Léopold Simar (2019). *Applied Multivariate Statistical Analysis* (5th ed.). Springer Nature. https://doi.org/10.1007/978-3-030-26006-4

[K] Inge Koch (2013). *Analysis of Multivariate and High-Dimensional Data*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025805

[V] Matthias Vallentin (Dec 19th, 2017). *Probability and Statistics Cookbook*. http://statistics.zone

[PP] Kaare B. Petersen, & Michael S. Pedersen (Nov 15th, 2012). *The Matrix Cookbook*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.113.6244

[Z] *Topics in Matrix Theory* (in Chinese). https://zhuanlan.zhihu.com/Topics-in-Matrix-Theory

List of Figures

List of Tables

Get your facts first, and then you can distort them as much as you please. — Mark Twain

§1 Review of Linear Algebra (2019/9/10)

For any $x, y \in \mathbb{R}^n$, their **inner product** is

$$\langle x, y\rangle = x'y = \sum_{k=1}^{n} x_k y_k.$$

The **projection** of x on $y \neq 0_n$ is

$$\mathrm{proj}_y\, x = \frac{\langle x, y\rangle}{\langle y, y\rangle} y.$$

Consider the linear model

$$Y = X\beta + \varepsilon,$$

where $X = (x_1, \ldots, x_p) \in \mathbb{R}^{n\times p}$ is called the **design matrix**. Usually, n is # of observations and p is # of variables. We are often faced with trade-off between *parsimony* and *goodness of fit*.

Choosing the maximal **linearly independent** system of x_k's leads to *dimension reduction*. Recall that $\mathrm{rank}\,(X)$ is the maximum # of linearly independent columns/rows, i.e., the dimension of the **column space** $\mathrm{Col}(X)$ or the **row space** $\mathrm{Col}(X')$.

Given a symmetric $A \in \mathbb{R}^{n\times n}$, the corresponding **quadratic form** is

$$x \in \mathbb{R}^n \mapsto x'Ax \in \mathbb{R}.$$

The matrix A is said to be **positive semi-definite** if $x'Ax \geq 0,\ \forall x \in \mathbb{R}^n$; furthermore, A is said to be **positive definite** if $x'Ax > 0$ whenever $x \neq 0_n$. Clearly, $X'X$ is non-negative definite.

The **eigen-decomposition** of symmetric $A \in \mathbb{R}^{n\times n}$ is

$$A = \sum_{i=1}^{n} \lambda_i e_i e_i' = (e_1, \ldots, e_n)\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)\,(e_1, \ldots, e_n)',$$

where λ_i's are eigenvalues and e_i's are orthonormal eigenvectors[i]. Note that $e_i e_i' = \mathrm{proj}_{e_i}$.

We say that $Q \in \mathbb{R}^{n\times n}$ is an **orthogonal matrix** if $Q'Q = I_n$. Any orthogonal matrix induces a map *preserving inner product*, i.e.,

$$\langle Qx, Qy\rangle = \langle x, y\rangle, \quad \forall x, y \in \mathbb{R}^n.$$

Write $A = Q\Lambda Q'$ with $Q = (e_1, \ldots, e_n)$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. If $y = Q'x$, then

$$x'Ax = y'\Lambda y = \sum_{k=1}^{n} \lambda_k y_k^2,$$

which is closely related to *ellipsoid* when A is positive definite. For any analytic function $f : \mathbb{R} \to \mathbb{R}$, we can define

$$f(A) := \sum_{i=1}^{n} f(\lambda_i) e_i e_i'.$$

Any orthogonal matrix of order 2 has the form

$$Q = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},\ \theta \in [0, 2\pi).$$

One can justify this fact by a few calculations using orthonormal relations.

§2 Review of Linear Algebra Cont'd (2019/9/17)

Given a positive definite $A \in \mathbb{R}^{2\times 2}$, write its spectral decomposition as $A = Q\Lambda Q'$, where

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2) = \begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix}, \qquad Q = (e_1, e_2) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

[i] cf. https://zhuanlan.zhihu.com/p/75250722

If $y = Q'x$, then $x'Ax = y'\Lambda y$. Note that

$$x = \begin{pmatrix} \|x\| \cos(\theta_x) \\ \|x\| \sin(\theta_x) \end{pmatrix} \implies y = \begin{pmatrix} \|x\| \cos(\theta_x - \theta) \\ \|x\| \sin(\theta_x - \theta) \end{pmatrix},$$

which is explained as a rotation. Henceforth,

$$\left\{ x \in \mathbb{R}^2 : x'Ax = c^2 \right\}$$

is an ellipse with the orientation defined by e_1, e_2 and length $\frac{c}{\sqrt{\lambda_1}}, \frac{c}{\sqrt{\lambda_2}}$.

Next, we explore **orthogonalization** for $X \in \mathbb{R}^{n \times p}$, i.e., to find $Z \in \mathbb{R}^{n \times p}$ with orthogonal columns such that $\mathrm{Col}(Z) = \mathrm{Col}(X)$. The so called **Gram-Schmidt process** makes it[ii]. To be explicit,

$$\begin{aligned} z_1 &= x_1, \qquad z_2 = x_2 - \mathrm{proj}_{z_1} x_2, \\ &\vdots \\ z_k &= x_k - \underbrace{\mathrm{proj}_{\mathrm{span}(z_1, \ldots, z_{k-1})} x_k}_{= \sum_{i=1}^{k-1} \mathrm{proj}_{z_i} x_k}, \qquad k \le p. \end{aligned}$$

It follows immediately that

$$x_k = \sum_{i=1}^{k} \mathrm{proj}_{z_i} x_k, \quad 1 \le k \le p,$$

which gives rise to $X = Z\Gamma$ with upper-triangular $\Gamma = (\gamma_{k\ell}) \in \mathbb{R}^{p \times p}$, where

$$\gamma_{k\ell} = \begin{cases} 1, & k = \ell; \\ \mathrm{proj}_{z_k} x_\ell, & k < \ell; \\ 0, & k > \ell. \end{cases}$$

Let $D = \mathrm{diag}(\|z_1\|, \ldots, \|z_p\|)$, then

$$Q = ZD^{-1}, \ R = D\Gamma \implies X = QR.$$

The **QR decomposition** involves $Q \in \mathbb{R}^{n \times p}$ with orthonormal columns and upper-triangular $R \in \mathbb{R}^{p \times p}$.

In linear regression (see LECTURE §13), the *hat matrix*

$$P_X = X(X'X)^{-1}X'$$

is actually the orthogonal projection onto $\mathrm{Col}(X)$. Using the QR decomposition, we obtain

$$P_X = QR(R'Q'QR)^{-1}R'Q' = QQ'.$$

Write $Q = (q_1, \ldots, q_p)$, then

$$\hat{y} = P_X y = \sum_{k=1}^{p} q_k q_k' y = \sum_{k=1}^{p} \langle y, q_k \rangle \, q_k.$$

The above results can be generalized for any separable Hilbert space, where we need apply some cut-off to pre-specified basis (e.g., Fourier basis, B-splines, wavelets, and other data-driven ones).

The **singular value decomposition** (SVD) is, say[iii],

$$X = U\Sigma V' = \sum_{k=1}^{\mathrm{rank}(X)} \sigma_k u_k v_k'$$

such that $Xv_k = \sigma_k u_k$ and $X'u_k = \sigma_k v_k$, where $\sigma_k > 0$ decreases in k. Note that

$$X'X = V\Sigma^2 V' = \sum_{k=1}^{\mathrm{rank}(X)} \sigma_k^2 v_k v_k' \implies P_X = UU' \implies \hat{y} = P_X y = \sum_{k=1}^{\mathrm{rank}(X)} \langle y, u_k \rangle \, u_k.$$

From the perspective of matrix completion, SVD provides the best *low-rank approximation* in that

$$\sum_{k=1}^{m} \sigma_k u_k v_k' = \underset{Y \in \mathbb{R}^{n \times p} : \mathrm{rank}(Y) \le m}{\arg\min} \mathrm{tr}\left((X - Y)(X - Y)' \right), \quad \forall m \le \mathrm{rank}\,(X)\,.$$

[ii] cf. https://zhuanlan.zhihu.com/p/75237479
[iii] cf. https://zhuanlan.zhihu.com/p/75283604

§3 Prediction (2019/9/19)

Early statistics focused on *inference*. Nowadays, people are more concerned about *prediction accuracy*. The AIC and BIC are frequently used for model selection[iv]. Generally, given data (X, Y), we are longing for

$$\min \ \mathbb{E}L(\hat{Y}, Y) = \mathbb{E}[\mathbb{E}[L(\hat{Y}, Y)|X]],$$

where L stands for loss function such as L^2 loss (risk) and 0-1 loss (cost).

- Let $\hat{Y} = f(X)$ and $L(\hat{y}, y) = |\hat{y} - y|^2$, then

$$\mathbb{E}[L(\hat{Y}, Y)|X] = \underbrace{\mathbb{E}\big[(Y - f(X))^2\big|X\big]}_{\text{MSE}} = \underbrace{\mathbb{E}\big[(Y - \mathbb{E}[Y|X])^2\big|X\big]}_{\text{Var}} + \underbrace{\big(\mathbb{E}[Y|X] - f(X)\big)}_{\text{Bias}}^2.$$

  The best predictor

$$\hat{f}(x) = \mathbb{E}[Y|X = x]$$

  for squared loss is appropriate for both fixed design and random/conditional design.

- In classification problem, $Y \in \{1, 2, \cdots, k\}$ is categorical and we often adopt 0-1 loss $L(\hat{y}, y) = \mathbb{1}_{[\hat{y} \neq y]}$. Let $\hat{Y} = g(X)$, then

$$\mathbb{E}[L(\hat{Y}, Y)|X = x] = \sum_{y \neq g(x)} \mathbb{P}(Y = y|X = x) = 1 - \mathbb{P}(Y = g(x)|X = x).$$

  Maximizing $\mathbb{P}(Y = g(x)|X = x)$ gives the **naïve Bayesian classifier** $\hat{g}(x)$ — the most probable class given x. We may treat y as a vector of dummy variables

$$y_j = \mathbb{1}_{[y=j]}, \quad j \in \{1, \cdots, k\},$$

  which add up to 1 and thus the degree of freedom is $k - 1$. Henceforth,

$$\hat{g}(x) = \underset{j \in \{1, \cdots, k\}}{\arg\max} \ \mathbb{E}[Y_j|X = x].$$

Most data fall into one of two groups: $\begin{cases} \text{numerical} \begin{cases} \text{continuous} \\ \text{discrete (counting)} \end{cases} \\ \text{categorical} \begin{cases} \text{nominal} \\ \text{ordinal} \end{cases} \end{cases}$

Given data $(x_i, y_i)_{1 \leq i \leq n}$, one may estimate

$$\mathbb{E}[Y|X = x] \approx \frac{1}{\#\{i : x_i = x\}} \sum_{x_i = x} y_i,$$

where $\{1 \leq i \leq n : x_i = x\}$ is the effective sample. If we replace it with the neighborhood $N_k(x)$, referred to as **k-nearest neighbor** (kNN), the kernel estimation (or kNN classifier) will come up. The case $k = 1$ results in *perfect*[v] *fit*, where no data reduction exists, and hereby the variance tends to explode even though the bias may vanish. The case when k is large will see larger bias and smaller variance. Anyway, we are desired to obtain suitable effective sample size, which could be rather difficult in real world.

§4 Prediction Cont'd & Multivariate Data (2019/9/24)

The *curse of dimensionality* turns out to be more significant when the model considered is nonparametric rather than parametric. The kNN estimator $\frac{1}{k}\sum_{x_i \in N_k(x)} y_i$ has $\int_{N_k(x)} f_X(u)\,\mathrm{d}u \propto f_X(x)h_x$ as its effective sample size, which often takes the form $\frac{A}{nh_x} + Bh_x^4$ of MSE = Var + Bias^2 due to the smoothness of the

[iv] cf. https://www.statlect.com/fundamentals-of-statistics/model-selection-criteria and Lecture §15
[v] ironic?

density function $f_X \in C^2$, where h_x is the neighborhood size (*bandwidth*). Trade-off between variance and bias provides the minimizer $h^* \propto n^{-1/5}$ and the minimum $\mathrm{MSE}(h^*) \propto n^{-4/5}$. Note that the parametric model has $\mathrm{MSE} \propto n^{-1}$. Generally, one may encounter

$$\mathrm{MSE}(h_x) = \frac{A}{nh_x^d} + Bh_x^b,$$

where d denotes the dimension and b is twice the order of differentiability of f_X. The optimal results are $h^* \propto n^{-1/(d+b)}$ and $\mathrm{MSE}(h^*) \propto n^{-b/(d+b)}$. For efficiency, we should always require $h \to 0$ and $nh^d \to \infty$ as the sample size n tends to infinity. See also exercise 5 in homework 1 for example.

The object to be studied is **data matrix**, which consists of observations and variables, denoted by

$$X = (x_{ij})_{1\le j\le p}^{1\le i\le n} = \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix} = (X_{(1)}, \ldots, X_{(p)}).$$

To check data quality, one may plot *histogram*, *boxplot*, *pairwise scatter plots*, etc. The concept of **outlier** is subtle, which suggests some alert. **Exploratory data analysis** helps finding structures in data — pattern, describe, visualize, PCA, cluster, etc. These are nonparametric (model-free) approaches, also called unsupervised learning. **Confirmatory data analysis** is model-based, including statistical inference and prediction. In other words, we assume

$$\textit{data} \sim \textit{model} + \textit{error}.$$

The model is wished both parsimony (as simple as possible, causing larger bias) and fidelity (adequate to describe data, causing larger variance). It's worth noting that[vi)]

> *all models are wrong, but some are useful.* — George E. P. Box

§5 Descriptive Statistics & Normal Distribution (2019/10/8)

Given data matrix $X = (x_{ij})_{1\le j\le p}^{1\le i\le n}$, the **sample mean** is defined to be

$$\bar{X} = (\bar{x}_1, \ldots, \bar{x}_p)', \quad \bar{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij} \text{ for } 1 \le j \le p.$$

The **sample covariance** is defined to be

$$S = (s_{jk})_{1\le j,k\le p}, \quad s_{jk} = \frac{1}{n-1}\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k).$$

The **sample correlation** is defined to be

$$R = (r_{jk})_{1\le j,k\le p}, \quad r_{jk} = s_{jk}/\sqrt{s_{jj}s_{kk}} \in [-1, 1].$$

Consider standardized data $x_{ij}^* = (x_{ij} - \bar{x}_j)/\sqrt{s_{jj}}$, its sample covariance is exactly $S^* = R = D^{-1}SD^{-1}$, where $D = \mathrm{diag}\big(\sqrt{s_{11}}, \ldots, \sqrt{s_{pp}}\big)$. The sample correlation characterizes the *linear* association between parameters. In linear regression with one regressor, the sample covariance appears naturally. Recall that the **Cauchy-Schwarz inequality** states that

$$\big(a'\sqrt{A}\sqrt{B}b\big)^2 \le (a'Aa)(b'Bb).$$

For *sampling statistics*, let $X_1, \ldots, X_n$ be i.i.d. random vectors with mean μ and covariance matrix Σ. The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ has mean $\frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i = \mu$ and covariance matrix $\frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{1}{n}\Sigma$. The denominator $(n-1)$ of the sample covariance

$$S = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'$$

[vi)] see https://www.zhihu.com/question/21416877/answer/806968779 for more fun.

guarantees the unbiasedness $\mathbb{E}S = \Sigma$. To see this, one can plug $\mathbb{E}X_iX_i' = \Sigma + \mu\mu'$ and $\mathbb{E}\bar{X}\bar{X}' = \frac{1}{n}\Sigma + \mu\mu'$ into the expectation of $(n-1)S = \sum_{i=1}^n X_iX_i' - n\bar{X}\bar{X}'$, and thus the corresponding **degree of freedom** (d.o.f.) is $n-1$ since one constraint $\sum_{i=1}^n (X_i - \bar{X}) = 0$ is imposed. Also, the MLE (see LECTURE §7)

$$\hat{\Sigma}_n = \frac{1}{n}(X_i - \bar{X})(X_i - \bar{X})'$$

for Σ in Gaussian setting is *biased*. In matrix language, $\bar{X} = \frac{1}{n}X'1_n$ and $(n-1)S = X'(I_n - \frac{1}{n}1_n1_n')X$. Denote the centering matrix by

$$H = I_n - \tfrac{1}{n}1_n1_n' \ \perp \ \tfrac{1}{n}1_n1_n'.$$

Clearly $H = H' = H^2$ is symmetric and idempotent, and thus $\operatorname{rank}(H) = \operatorname{tr}(H) = n-1$.

Probability theory deduces the underlying structure of observed data/samples. Statistical inference aims to learn unknown (or ad hoc) population/distribution/parameters from the observations.

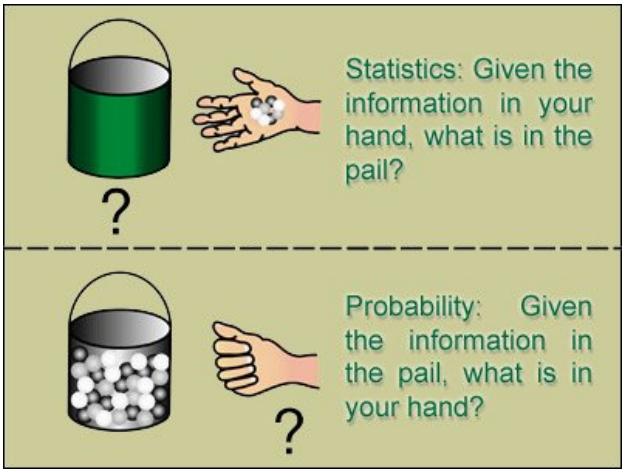

Figure 1: Difference between statistics and probability[vii)]

In multivariate framework, we often specify notations in this manner:

- $\mathbb{E}X = \mu = (\mu_j)_{1\le j\le p}$;
- $\operatorname{Var}(X) = \Sigma = (\sigma_{jk})_{1\le j,k\le p}$;
- $\boldsymbol{\rho} = (\rho_{jk})_{1\le j,k\le p}$ with $\rho_{jk} = \sigma_{jk}/\sqrt{\sigma_{jj}\sigma_{kk}}$; and
- $\operatorname{Cov}(X,Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)' = \Sigma_{XY}$. Note that $\operatorname{Cov}(AX + a, BY + b) = A\Sigma_{XY}B'$.

To justify the **normal distribution**, the *central limit theorem* (CLT) states that if $X_1, X_2, \ldots$ are i.i.d. random variables with mean μ and variance σ^2, then $\sqrt{n}(\bar{X}_n - \mu)/\sigma \to \mathcal{N}(0,1)$ in distribution. For non-degenerate multivariate normal distribution $\mathcal{N}_p(\mu, \Sigma)$, the probability density function (p.d.f.) is

$$f(x) = \frac{1}{(2\pi)^{p/2}(\det\Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}, \ x \in \mathbb{R}^p.$$

The crucial part is Σ^{-1}, called the **precision matrix**, which is important in graphical models. Generally, independence implies uncorrelatedness, but not vice versa. If random p-vector X_1 and random q-vector X_2 are jointly normal, then $\operatorname{Cov}(X_1, X_2) = 0$ if and only if $X_1 \perp\!\!\!\perp X_2$. The conditional distribution is more complicated, involving some matrix holing tricks[viii)]. Some frequently used properties are as follows:

i) $X \sim \mathcal{N}_p(\mu, \Sigma) \implies AX + b \sim \mathcal{N}_q(A\mu + b, A\Sigma A')$ for $A \in \mathbb{R}^{q\times p}$ and $b \in \mathbb{R}^q$;

ii) if $a'X \sim \mathcal{N}(a'\mu, a'\Sigma a)$ for any $a \in \mathbb{R}^p$, then $X \sim \mathcal{N}_p(\mu, \Sigma)$;

iii) $X \sim \mathcal{N}_p(\mu, \Sigma) \iff \mathbb{E}\exp(it'X) = \exp\left(it'\mu - \frac{1}{2}t'\Sigma t\right)$, $t \in \mathbb{R}^p$, where $i = \sqrt{-1}$.

[vii)] from Herman Bennett's *14.30 Introduction to Statistical Method in Economics*. Spring 2006. https://ocw.mit.edu

[viii)] cf. https://zhuanlan.zhihu.com/p/78884647

§6 Multivariate Normal Distribution Cont'd & MLE (2019/10/15)

The distribution of any random vector X corresponds uniquely to all one-dimensional distributions of its linear functions $a'X$. To see this, the **characteristic function** $\phi_X(t) = \mathbb{E}\exp(\sqrt{-1}t'X)$ applies.

For inference, we consider $\mathbb{P}\left\{\frac{|X-\mu|}{\sigma} \le c\right\}$ for random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ and constant $c \in \mathbb{R}_+$. If instead $X \sim \mathcal{N}_p(\mu, \Sigma)$ is a random vector, the region considered becomes $\{(X-\mu)'\Sigma^{-1}(X-\mu) \le c^2\}$, which correspond to **contours with constant density** $\{x \in \mathbb{R}^p : f_X(x) \ge c_0^2\}$. Write $\Sigma = Q\Lambda Q'$ with $Q = (e_1, \dots, e_p)$ orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$. Let $Y = (Y_1, \dots, Y_p)' = Q'(X-\mu) \sim \mathcal{N}_p(0_p, \Lambda)$, then

$$(X-\mu)'\Sigma^{-1}(X-\mu) = Y'\Lambda^{-1}Y = \sum Y_j^2/\lambda_j \sim \chi_p^2,$$

geometrically calling on some ellipsoid.

Suppose $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$, then $X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$ is independent of X_2, and thus

$$X_1|X_2 \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

Note that $\mathbb{E}[X_1|X_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2)$ minimizes $\mathbb{E}|Z-X_1|^2$ among all $Z = g(X_2)$, i.e., the conditional expectation is the population version of the *best linear prediction*, compared to the sample version $\hat{Y} = \bar{Y} + S_{YX}S_{XX}^{-1}(X-\bar{X})$ in linear regression[ix] (see LECTURE §13).

We now explore the *maximum likelihood estimation* (MLE) for $\theta \in \Theta$ given a sample $x_1, \dots, x_n$. The likelihood function is the joint density

$$L_n(\theta|x_1, \dots, x_n) = f(x_1, \dots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta).$$

Hence, the log-likelihood function is

$$\ell_n(\theta) = \log L_n(\theta|x_1, \dots, x_n) = \sum_{i=1}^n \log f(x_i|\theta).$$

To justify the MLE, suppose X_i's are i.i.d. with p.d.f. $f_0 = f(\cdot|\theta_0)$. **Shannon's entropy inequality** states that

$$\mathbb{E}\log f(X) \le \mathbb{E}\log f_0(X)$$

for any p.d.f. $f = f(\cdot|\theta)$, with equality holding if and only if $f = f_0$ almost everywhere. Note that the **identifiablity** means

$$\theta_1 \ne \theta_2 \iff f(\cdot|\theta_1) \ne f(\cdot|\theta_2).$$

By SLLN, $\frac{1}{n}\ell_n(\theta) \to \mathbb{E}\log f(X|\theta)$ almost surely as $n \to \infty$, so the maximizer $\hat{\theta}_n^{\mathrm{MLE}}$ of $\ell_n(\cdot)$ seems reasonable to estimate the true parameter θ_0. Under some mild assumptions, $\hat{\theta}_n^{\mathrm{MLE}}$ is consistent, as desired.

§7 Maximum Likelihood Estimation Cont'd (2019/10/17)

In multivariate normal setting,

$$\begin{aligned} L_n(\mu, \Sigma) &= \prod_{i=1}^n \left[\frac{1}{(2\pi)^{p/2}(\det\Sigma)^{1/2}}\exp\left\{-\frac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right\}\right] \\ &= \frac{1}{(2\pi)^{np/2}(\det\Sigma)^{n/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^n(x_i-\bar{x})'\Sigma^{-1}(x_i-\bar{x})\right\}\exp\left\{-\frac{n}{2}(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)\right\}, \end{aligned}$$

so $\hat{\mu}_{\mathrm{MLE}} = \bar{x}$ and $\hat{\Sigma}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^n(x_i-\bar{x})(x_i-\bar{x})'$ (exercise 2 in homework 2) with $\mathrm{Bias}(\hat{\Sigma}_{\mathrm{MLE}}) = -\frac{1}{n}\Sigma$. They are asymptotically consistent, minimal sufficient and complete[x].

[ix] *regression toward the mean* is the phenomenon that arises if a random variable is extreme on its first measurement but closer to the mean or average on its second measurement and if it is extreme on its second measurement but closer to the average on its first.

[x] a statistic $T = T(X)$ is said to be **complete** for $\theta \in \Theta$ if, $\mathbb{E}_\theta f(T) = 0$ for all $\theta \in \Theta$ implies $f(T) = 0$ a.s., and thus T appears to be successful in reducing the data.

To derive MLEs, we can use matrix differentiation[xi)] as well. For symmetric positive definite $A = (a_{ij})$,

- $\partial(x'Ax)/\partial x = 2Ax$;
- $\partial \operatorname{tr}(AB)/\partial A = 2B - \operatorname{diag}(B)$;
- $\partial \log\det(A)/\partial a_{ij} = \begin{cases} 2A_{ij}/\det(A), & i \neq j \\ A_{ii}/\det(A), & i = j \end{cases}$ and thus $\partial \log\det(A)/\partial A = 2A^{-1} - \operatorname{diag}\big(A^{-1}\big)$.

Note that $\bar{X} \sim \mathcal{N}_p(\mu, \frac{1}{n}\Sigma)$ is independent of $n\hat{\Sigma}_{\mathrm{MLE}} \sim W_p(\Sigma, n-1)$. See Lecture §9 for the *Wishart distribution*. As the sample size $n \to \infty$, the following large sample properties hold:

- $\bar{X} \to_{\mathbb{P}} \mu$, $\hat{\Sigma} \to_{\mathbb{P}} \Sigma$;
- $\sqrt{n}(\bar{X} - \mu) \to_d \mathcal{N}_p(0_p, \Sigma)$;
- $n(\bar{X} - \mu)'\hat{\Sigma}^{-1}(\bar{X} - \mu) \to_d \chi^2_p$.

Generally, the derivative of $\ell_n(\theta)$ is called the **score function**, denoted by $S_n(\theta)$. Under some regularity conditions,

$$\mathbb{E}_\theta S(\theta) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = \int \frac{1}{\cancel{f(x|\theta)}}\frac{\partial f(x|\theta)}{\partial\theta}\cancel{f(x|\theta)}\,\mathrm{d}x = \frac{\partial}{\partial\theta}\underbrace{\int f(x|\theta)\,\mathrm{d}x}_{=1} = 0_p,$$

and (check!)

$$-\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta'}S_n(\theta)\right] = \mathbb{E}_\theta[S_n(\theta)S_n(\theta)'] = \operatorname{Var}_\theta(S_n(\theta))$$

is exactly the **Fisher information** $\mathcal{I}_n(\theta)$. Note that $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$. As $n \to \infty$,

- $\sqrt{n}(\hat{\theta}_{\mathrm{MLE}} - \theta) \to_d \mathcal{N}_p\big(0_p, \mathcal{I}_1(\theta)^{-1}\big)$;
- $n(\hat{\theta}_{\mathrm{MLE}} - \theta)'\mathcal{I}_1(\hat{\theta}_{\mathrm{MLE}})(\hat{\theta}_{\mathrm{MLE}} - \theta) \to_d \chi^2_p$.

Sketch of Proof. Denote by $H_n(\theta) = \frac{\partial}{\partial\theta'}S_n(\theta)$ the Hessian matrix of $\ell_n(\theta)$. Then

$$0_p = S_n(\hat{\theta}_{\mathrm{MLE}}) \approx S_n(\theta) + H_n(\theta)(\hat{\theta}_{\mathrm{MLE}} - \theta) \implies \sqrt{n}(\hat{\theta}_{\mathrm{MLE}} - \theta) \approx \big(-\tfrac{1}{n}H_n(\theta)\big)^{-1}\tfrac{1}{\sqrt{n}}S_n(\theta),$$

where $-\frac{1}{n}H_n(\theta) \to_{\mathbb{P}} \mathcal{I}_1(\theta)$ and $\frac{1}{\sqrt{n}}S_n(\theta) \to_d \mathcal{N}_p(0_p, \mathcal{I}_1(\theta))$. □

§8 Diagnostics for Normality & Matrix Distribution (2019/10/22)

For univariate data, ordered as $x_{(1)} \le \cdots \le x_{(n)}$ called *sample quantiles*, the most useful diagnostic tool is the **Q-Q plot**. Let $q_{(1)}, \ldots, q_{(n)}$ be *theoretical quantiles* of $\mathcal{N}(0,1)$, i.e.,

$$\Phi(q_{(i)}) = \frac{i - \frac{1}{2}}{n}, \quad 1 \le i \le n.$$

Plot $x_{(i)}$ against $q_{(i)}$, and then compare it to a line. If $X_i \sim \mathcal{N}_p(\mu, \Sigma)$ are multivariate, then

$$(X_i - \mu)'\Sigma^{-1}(X_i - \mu) \sim \chi^2_p.$$

By large sample property $\bar{X} \to_{\mathbb{P}} \mu$ and $S \to_{\mathbb{P}} \Sigma$, as $n \gg p$,

$$d_i^2 := (X_i - \bar{X})'S^{-1}(X_i - \bar{X}) \overset{\circ}{\sim} \chi^2_p.$$

Sort d_i^2 to obtain *sample quantiles* $d^2_{(i)}$, and plot them against *theoretical quantiles* of χ^2_p.

[xi)] cf. https://cs.nju.edu.cn/wujx/teaching/MatrixCookBook.pdf

To detect outliers and clean data, one may plot histogram or dot plot for each variable, and pairwise scatterplot for each pair. Sometimes we prefer **standardized variable**

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_{kk}}, \quad 1 \le i \le n,\ 1 \le k \le p.$$

Note that judging outliers should take both n and p into consideration. The empirical 3σ *rule*[xii] is not always appropriate.

Transforms are used to normalize univariate data, including

- $y \in \mathbb{N} \mapsto \sqrt{y}$ for counts;
- $\text{logit} : y \in [0,1] \mapsto \log\left(\frac{y}{1-y}\right)$ for proportions;
- $r \in [-1,1] \mapsto \log\left(\frac{1+r}{1-r}\right)$ for correlation coefficients;
- **Box-Cox transform** $x \mapsto x^{(\lambda)} = \begin{cases} (x^\lambda - 1)/\lambda, & \lambda > 0 \\ \log(x), & \lambda = 0 \end{cases}$, where λ is a tuning parameter to maximize $\ell(\lambda) = -\frac{n}{2}\log\left(\frac{1}{n}\sum_{i=1}^n \left|x_i^{(\lambda)} - \overline{x^{(\lambda)}}\right|^2\right) + (\lambda - 1)\sum_{i=1}^n \log(x_i)$, or equivalently, to minimize the sample variance of $y_i^{(\lambda)} := \frac{x_i^\lambda - 1}{\lambda(\prod_{j=1}^n x_j)^{(\lambda-1)/n}}$. For multivariate data, we often choose λ_k for X_{ik} separately, pretending the components are independent.

Next, multivariate data should be tackled. If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$, then it is said that

$$X = \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix} = (X_{(1)}, \ldots, X_{(p)}) \sim \mathcal{MN}_{n\times p}(1_n\mu', I_n, \Sigma)$$

is a **normal matrix** with p.d.f.

$$f(X) = \det(2\pi\Sigma)^{-n/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(X - 1_n\mu')'(X - 1_n\mu')\right)\right\}.$$

Using vectorization and Kronecker product[xiii], we have

$$\operatorname{vec}(X) = \begin{pmatrix} X_{(1)} \\ \vdots \\ X_{(p)} \end{pmatrix} \sim \mathcal{N}_{np}(\mu \otimes 1_n, \Sigma \otimes I_n).$$

Several frequently used properties are linearity, mixed-product property, and vectorization formula

$$\operatorname{vex}(AXB) = (B' \otimes A)\operatorname{vec}(X).$$

It follows that (see exercise 4 in homework 2)

$$\operatorname{vex}(AXB) \sim \mathcal{N}_{mq}((B'\mu) \otimes (A1_n), (B'\Sigma B) \otimes (AA'))$$

for $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$. If $\mu = 0_p$ and $Q \in \mathbb{R}^{n\times n}$ is orthogonal, then

$$\operatorname{vec}(QX) =_d \operatorname{vec}(X) \sim \mathcal{N}_{np}(0_p \otimes 1_n, \Sigma \otimes I_n).$$

Let $Y = AXB$ and $Z = CXD$, then

$$\operatorname{Cov}(\operatorname{vec}(Y), \operatorname{vec}(Z)) = (B'\Sigma D) \otimes (AC'),$$

so $Y \perp\!\!\!\perp Z$ if and only if $AC' = 0$ or $B'\Sigma D = 0$. As an application, to see $\hat{\mu} = \frac{1}{n}X'1_n$ independent of $\hat{\Sigma} = \frac{1}{n}X'(I_n - \frac{1}{n}1_n1_n')X$, denoting $H = I_n - \frac{1}{n}1_n1_n'$, it suffices that

$$\operatorname{Cov}(\operatorname{vec}(1_n'X), \operatorname{vec}(HX)) = \Sigma \otimes (1_n'H) = 0.$$

[xii] cf. http://wikipedia.moesalih.com/Three-sigma_rule
[xiii] cf. https://zhuanlan.zhihu.com/p/75499831

§9 Cochran's Theorem & Hypothesis Testing (2019/10/29)

The celebrated **Cochran's theorem**[xiv] concerns *simultaneous diagonalization*[xv] to idempotent matrices.

> Let $X = (X_1, \ldots, X_n)' \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$. Suppose $\sum_{i=1}^n X_i^2 = \sum_{j=1}^k Q_j$ with $Q_j = X' A_j X$ for some symmetric $A_j \in \mathbb{R}^{n\times n}$ of rank r_j. That is, $I_n = \sum_{j=1}^k A_j$. If $\sum_{j=1}^k r_j = n$, then there exists some orthogonal $C \in \mathbb{R}^{n\times n}$ such that $C'X = (Y_1, \ldots, Y_n)'$ satisfies
>
> $$Q_j = \sum_{i=1+r_1+\cdots+r_{j-1}}^{r_1+\cdots+r_{j-1}+r_j} Y_i^2, \qquad 1 \le j \le k.$$
>
> Note that $Q_j \sim \sigma^2 \chi^2_{r_j}$ $(j = 1, \cdots, k)$ are independent.

For a multivariate version, some acquaintance of the **Wishart distribution** is prerequisite. Consider the data matrix $X = (X_1, \ldots, X_n)'$ with $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0_p, \Sigma)$, it is said that $X'X = \sum_{i=1}^n X_i X_i' \sim W_p(\Sigma, n)$.

- $W_1(\sigma^2, n) = \sigma^2 \chi^2_n$ when $p = 1$;
- $BMB' \sim W_q(B\Sigma B', n)$ for $M \sim W_p(\Sigma, n)$ and $B \in \mathbb{R}^{q\times p}$;
- $a'Ma \sim (a'\Sigma a)\chi^2_n$ for $M \sim W_p(\Sigma, n)$ and $a \in \mathbb{R}^p$;
- $M_1 + M_2 \sim W_p(\Sigma, n_1 + n_2)$ for independent $M_i \sim W_p(\Sigma, n_i)$, $i = 1, 2$.

Let $X = (X_1, \ldots, X_n)'$ with $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0_p, \Sigma)$, and $A_1, \ldots, A_k, A, B \in \mathbb{R}^{n\times n}$ be symmetric, then

(a) $X'AX$ is a $\lambda(A)$-weighted sum of independent $W_p(\Sigma, 1)$ matrices;

(b) $X'AX \sim W_p(\Sigma, r)$ if and only if A is idempotent of rank $r \ge p$;

(c) $X'AX \perp\!\!\!\perp X'BX$ if and only if $AB = 0$.

If $X'X = \sum_{j=1}^k X'A_jX$ (i.e., $I_n = \sum_{j=1}^k A_j$) and $\sum_{j=1}^k r_j = n$, where $r_j = \operatorname{rank}(A_j)$ for $1 \le j \le k$, then $X'A_jX$ are independent and distributed as $W_p(\Sigma, r_j)$ matrices, respectively.

Since $H = I_n - \frac{1}{n}1_n 1_n'$ has a spectral decomposition $H = C\Lambda C'$ such that $\Lambda = \operatorname{diag}(I_{n-1}, 0)$ and $C = (e_1, \ldots, e_{n-1}, e_n)$ with $e_n = \frac{1}{\sqrt{n}}1_n$, denoting $Y = (Y_1, \ldots, Y_n)' = C'X$, it follows that

$$n\hat{\Sigma} = (n-1)S = X'HX = Y'\Lambda Y = \sum_{i=1}^{n-1} Y_i Y_i' \sim W_p(\Sigma, n-1).$$

From the perspective of Cochran's theorem,

$$X'X = X'(I_n - \tfrac{1}{n}1_n 1_n')X + X'(\tfrac{1}{n}1_n 1_n')X,$$

where $I_n - \frac{1}{n}1_n 1_n'$ and $\frac{1}{n}1_n 1_n'$ are idempotent and of rank $n-1$ and 1, respectively.

Next, we explore inferences for the mean of normal distribution. For univariate $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\sigma > 0$ unknown, $T = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$ is often employed to test $H_0 : \mu = \mu_0$, since $T \overset{H_0}{\sim} t_{n-1}$ no matter which value μ_0 takes. To reject H_0 at level α, one can adopt either the criterion $\{|T| > t_{n-1}(\alpha/2)\}$, or *equivalently*, that $\mu_0 \notin \left[\bar{X} - \frac{s}{\sqrt{n}}t_{n-1}(\alpha/2), \bar{X} + \frac{s}{\sqrt{n}}t_{n-1}(\alpha/2)\right]$. Note that $T^2 = n(\bar{X} - \mu_0)(s^2)^{-1}(\bar{X} - \mu_0) \overset{H_0}{\sim} F_{1,n-1}$.

In multivariate setting, if $\mathbf{d} \sim \mathcal{N}_p(0_p, I_p)$ and $\mathbf{M} \sim W_p(I_p, m)$ are independent, then

$$m\mathbf{d}'\mathbf{M}^{-1}\mathbf{d} \sim T^2(p, m) = \frac{mp}{m-p+1}F_{p,m-p+1},$$

called **Hotelling's T-squared distribution**. Let $\mathbf{X}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$. Now that $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ and $(n-1)\mathbf{S} \sim W_p(\Sigma, n-1)$ are independent, it's immediate that

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) = (n-1)\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})'\left((n-1)\mathbf{S}\right)^{-1}\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim T^2(p, n-1).$$

Hotelling's T^2 statistic is clearly a generalization of *Student's t-statistic* for testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$. Hence, a $100(1-\alpha)\%$ confidence region/*ellipsoid* for $\boldsymbol{\mu}$ is $\left\{\boldsymbol{\mu} : n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \le \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right\}$. For a large sample ($n \gg p$), we have $n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \overset{\circ}{\sim} \chi^2_p$, so $\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$ can be replaced by $\chi^2_p(\alpha)$.

[xiv] cf. https://zhuanlan.zhihu.com/p/85314314
[xv] cf. https://zhuanlan.zhihu.com/p/75250722

§10 Hypothesis Testing Cont'd (2019/10/31)

Generally, the **likelihood-ratio test** statistic for $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta \setminus \Theta_0$ is

$$\Lambda = \frac{\max_{\theta\in\Theta_0} L(\theta)}{\max_{\theta\in\Theta} L(\theta)},$$

with d.o.f.$= \dim(\Theta) - \dim(\Theta_0)$. Wilks' theorem asserts that

$$-2\log(\Lambda) \to_d \chi^2_{\dim(\Theta)-\dim(\Theta_0)}.$$

The rejection region has the form $\{\Lambda < c\}$ for some c.

Consider $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$ and $H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$. The likelihood function is

$$L(\mu, \Sigma) = (2\pi)^{-np/2} \det(\Sigma)^{-n/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(X - 1_n\mu')'(X - 1_n\mu')\right)\right\}.$$

Hence (see exercise 2 in homework 2)

$$\max_{\mu,\Sigma} L(\mu, \Sigma) = (2\pi)^{-np/2} \det(\hat{\Sigma})^{-n/2} e^{-np/2},$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'$, and

$$\max_{\Sigma} L(\mu_0, \Sigma) = (2\pi)^{-np/2} \det(\hat{\Sigma}_0)^{-n/2} e^{-np/2},$$

where $\hat{\Sigma}_0 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)(X_i - \mu_0)'$ is the constrained/restricted MLE. It follows that

$$\Lambda^{2/n} = \frac{\det(\hat{\Sigma})}{\det(\hat{\Sigma}_0)} = \frac{\det\left(\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'\right)}{\det\left(\sum_{i=1}^n (X_i - \mu_0)(X_i - \mu_0)'\right)}$$

is equivalent to Hotelling's T^2 in the sense of monotone bijection, since

$$\sum_{i=1}^n (X_i - \mu_0)(X_i - \mu_0)' = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})' + n(\bar{X} - \mu_0)(\bar{X} - \mu_0)',$$

the *decomposition of the sums of squares*, implies (see exercise 1 in homework 1)

$$\Lambda^{-2/n} = 1 + n(\bar{X} - \mu_0)' \left((n-1)S\right)^{-1} (\bar{X} - \mu_0) = 1 + \frac{1}{n-1} T^2.$$

To test $H_0 : \theta = \theta_0$ in other ways, denoting by $\hat{\theta} = \arg\max_{\theta\in\Theta} L(\theta)$ the unconstrained MLE,

- the **Wald test** employs

  $$W = (\hat{\theta} - \theta_0)' \widehat{\operatorname{Var}}(\hat{\theta})^{-1} (\hat{\theta} - \theta_0) = (\hat{\theta} - \theta_0)' \mathcal{I}_n(\hat{\theta}) (\hat{\theta} - \theta_0) \xrightarrow[H_0]{d} \chi^2_{\dim(\Theta)},$$

  noting that $\operatorname{Var}_\theta(\hat{\theta}) \approx \mathcal{I}_n(\theta)^{-1}$ (see LECTURE §7); and

- the **score test** (a.k.a. the **Lagrange multiplier test**, particularly in econometrics) uses

  $$U_n(\theta_0)' \mathcal{I}_n(\theta_0)^{-1} U_n(\theta_0) \xrightarrow[H_0]{d} \chi^2_{\dim(\Theta)},$$

  where $U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta}$ is the score.

The three tests above are equivalent for large samples. An advantage of the Wald test over the other two is that it only requires the estimation of the unrestricted model, which lowers the computational burden. The score test only requires the estimation of the likelihood function under the null hypothesis, and thus it is less specific than the other two tests about the precise nature of the alternative hypothesis[xvi].

[xvi] a.k.a. the *research hypothesis*, being more important and meaningful than the null hypothesis (*benchmark*).

The **multiple testing problem** occurs when one considers a set of statistical inferences *simultaneously*. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$ with Σ unknown. For any single $a \in \mathbb{R}^p$,

$$T(a) = \frac{\sqrt{n}(a'\bar{X} - a'\mu)}{\sqrt{a'Sa}} \sim t_{n-1}$$

gives a $100(1-\alpha)\%$ confidence interval $a'\bar{X} \pm t_{n-1}(\frac{\alpha}{2})\sqrt{\frac{a'Sa}{n}}$ for $a'\mu$. To obtain some $c > 0$ such that

$$\mathbb{P}\left\{|T(a)|^2 \le c^2,\ \forall a \in \mathbb{R}^p\right\} = 1 - \alpha,$$

note that

$$\max_a |T(a)|^2 = \max_a \frac{n\left|a'(\bar{X} - \mu)\right|^2}{a'Sa} = n(\bar{X} - \mu)'S^{-1}(\bar{X} - \mu) = T^2$$

by Cauchy-Schwarz inequality, so

$$\mathbb{P}\left(\bigcap_{a \in \mathbb{R}^p} \left\{a'\mu \in a'\bar{X} \pm \sqrt{\tfrac{(n-1)p}{n-p} F_{p,n-p}(\alpha) \tfrac{a'Sa}{n}}\right\}\right) \ge 1 - \alpha.$$

In other words, replacing $t_{n-1}(\frac{\alpha}{2}) = \sqrt{F_{1,n-1}(\alpha)}$ with $\sqrt{\frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)}$ yields *simultaneous confidence intervals* for all possible linear functions of μ.

When a small number of statements are of interest, the **Bonferroni correction** can be used to adjust confidence intervals. If $\mathbb{P}(C_k) = 1 - \alpha_k$ for $1 \le k \le m$, then

$$\mathbb{P}(\cap C_k) = 1 - \mathbb{P}(\cup C_k^{\complement}) \ge 1 - \sum \alpha_k,$$

and thus $\mathbb{P}(\cap C_k) \ge 1 - \alpha$ as long as $\alpha_k = \alpha/m$. For example,

$$\mathbb{P}\left(\bigcap_{k=1}^{m} \left\{a_k'\mu \in a_k'\bar{X} \pm t_{n-1}(\tfrac{\alpha}{2m})\sqrt{\tfrac{a'Sa}{n}}\right\}\right) \ge 1 - \alpha, \quad \forall a_1, \dots, a_m \in \mathbb{R}^p.$$

§11 Inferences about Multi-Population Means (2019/11/5)

To compare two treatments, the same units $i = 1, 2, \cdots, n$ are assigned to both so that the comparison makes sense. Formally, consider $X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{2p}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$, and $H_0 : \mu_1 = \mu_2$ is of interest. Let $D_i = X_{1i} - X_{2i}$ and $\mu_d = \mu_1 - \mu_2$. It suffices to test $H_0 : \mu_d = 0_p$ using

$$T_D^2 = n\bar{D}'S_D^{-1}\bar{D} \overset{H_0}{\sim} T^2(p, n-1) = \frac{(n-1)p}{n-p} F_{p,n-p}.$$

If $C \in \mathbb{R}^{q \times p}$ has full row rank $q \le p$, then the statistic used to test the constraint

$$H_0 : C\mu = \nu_0 \ (= 0_q)$$

on the population mean $\mu \in \mathbb{R}^p$ of $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$ is

$$T^2 = n(C\bar{X} - \nu_0)'(CSC')^{-1}(C\bar{X} - \nu_0) \overset{H_0}{\sim} \frac{(n-1)q}{n-q} F_{q,n-q},$$

since $CX_i \sim \mathcal{N}_q(C\mu, C\Sigma C')$. Note that T^2 does not depend on the particular choice of C, i.e., T^2 computed with $\tilde{C}$ instead of C yields the same result provided that $\tilde{C} = QC$ for some invertible $Q \in \mathbb{R}^{q \times q}$.

A **contrast** is a linear combination of variables whose coefficients add up to zero. Attention is usually centered on contrasts when comparing different treatments. The example at the beginning introduces $C = (I_p, -I_p)$ as a *contrast matrix*[xvii], from which $D_i = CX_i$, $\bar{D} = C\bar{X}$ and $S_D = CS_XC'$ are derived.

[xvii] cf. https://stats.stackexchange.com/a/221861

In *repeated measurements*, subjects i are assumed independent, but treatments j within each subject are dependent, which differs from that in one-way univariate ANOVA (see also LECTURE §12). The responses are

$$x_{ij} = \mu_j + \varepsilon_{ij}, \quad i = 1, \cdots, n;\ j = 1, \cdots, p.$$

Here $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ip})' \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0_p, \Sigma)$ are random errors. In order to test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$, one may take

$$C = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix} \text{ or } \begin{pmatrix} 1 & -1 & & & \\ 1 & & -1 & & \\ \vdots & & & \ddots & \\ 1 & & & & -1 \end{pmatrix} \in \mathbb{R}^{(p-1)\times p},$$

each of which gives

$$T^2 = n(C\bar{X})'(CSC')^{-1}(C\bar{X}) \overset{H_0}{\sim} \frac{(n-1)(p-1)}{n-p+1} F_{p-1,n-p+1}.$$

If we write $\mu_j = \mu + \alpha_j$, then $(\mu, \alpha_1, \ldots, \alpha_p)' \in \mathbb{R}^{p+1}$ is not identifiable/estimable, but $H_0 : \alpha_1 = \cdots = \alpha_p$ is indeed estimable and thus can be tested. It's often imposed that $\sum_{j=1}^p \alpha_j = 0$.

Next, we will compare means from *two populations*. Suppose $X_{11}, \ldots, X_{1n_1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu_1, \Sigma_1)$ and $X_{21}, \ldots, X_{2n_2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu_2, \Sigma_2)$ are two independent samples. The goal is to test $H_0 : \mu_1 = \mu_2$. Assume $\Sigma_1 = \Sigma_2$. Note that the first-order information is meaningful only if the second-order information keeps almost the same. Denote

$$\bar{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ki}$$

and

$$S_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)'$$

for $k = 1, 2$. Let

$$S_{\text{pooled}} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2},$$

whose d.o.f. is $n_1 + n_2 - 2$ since two sample means are computed separately. It holds that

$$\left(\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)\right)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{\text{pooled}}\right]^{-1} \left(\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)\right) \sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p,n_1+n_2-p-1}.$$

For large samples, having no need of $\Sigma_1 = \Sigma_2$ or normality,

$$\left(\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)\right)' \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right)^{-1} \left(\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)\right) \overset{\circ}{\sim} \chi^2_p.$$

Profile analysis pertains to situations in which a series of treatments are administered to two or more groups of subjects. It is assumed that (i) all measurements are expressed in similar scales; and (ii) all subjects are independent. Three questions on the population means $\mu_1, \mu_2 \in \mathbb{R}^p$ are formulated stepwise:

(1) Are the profiles parallel in the sense that $\mu_1 - \mu_2$ is constant?

(2) If (1) holds, are the profiles the same in the sense that $\mu_1 = \mu_2$?

(3) If (2) holds, are the treatments identical in the sense that $\mu_{,1} = \cdots = \mu_{,p}$?

To that end, denoting the contrast matrix by

$$C = \begin{pmatrix} -1 & & & 1 \\ & -1 & & 1 \\ & & \ddots & \vdots \\ & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(p-1)\times p},$$

we can carry out the sequential tests as follows:

(1) Test $H_{01}: C\mu_1 = C\mu_2$ using

$$\left(\bar{X}_1 - \bar{X}_2\right)' C' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) C S_{\text{pooled}} C'\right]^{-1} C\left(\bar{X}_1 - \bar{X}_2\right) \overset{H_0}{\sim} \frac{(n_1+n_2-2)(p-1)}{n_1+n_2-p} F_{p-1,n_1+n_2-p}.$$

If H_{01} is rejected, stop; otherwise, continue.

(2) Test $H_{02}: 1_p'\mu_1 = 1_p'\mu_2$ using

$$1_p'\left(\bar{X}_1 - \bar{X}_2\right) \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) 1_p' S_{\text{pooled}} 1_p\right]^{-1} 1_p'\left(\bar{X}_1 - \bar{X}_2\right) \overset{H_0}{\sim} F_{1,n_1+n_2-2} = t^2_{n_1+n_2-2}.$$

If H_{02} is rejected, stop; otherwise, continue.

(3) Test $H_{03}: C\mu = 0_{p-1}$ using

$$(n_1+n_2)\bar{X}'C'(CSC')^{-1}C\bar{X} \overset{H_0}{\sim} \frac{(n_1+n_2-1)(p-1)}{n_1+n_2-p+1} F_{p-1,n_1+n_2-p+1},$$

where $\bar{X}$ and S are based on all n_1+n_2 observations.

§12 Multivariate Analysis of Variance (2019/11/12)

Usually, more than two treatments (e.g. drugs) are administered. In the *one-way* (single-factor) M**ANOVA** (multivariate analysis of variance) model,

$$\underset{\text{(observation)}}{X_{\ell r}} = \underset{\text{(overall mean)}}{\mu} + \underset{\text{(the } \ell_{\text{th}} \text{ treatment effect)}}{\tau_\ell} + \underset{\text{(error)}}{\varepsilon_{\ell r}} \quad (r = 1, \cdots, n_\ell;\ \ell = 1, \cdots, g)$$

represents the p-dimensional response under the ℓ_{th} treatment in the r_{th} replication, where the constraint $\sum_{\ell=1}^g n_\ell \tau_\ell = 0_p$ is specified. The errors are assumed to be $\varepsilon_{\ell r} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0_p, \Sigma)$, which only influences the distribution of the test statistic.

ANOVA is a special case of linear models (see LECTURE §13), and here

$$\begin{pmatrix} X_{11} \\ \vdots \\ X_{1n_1} \\ \vdots \\ X_{g1} \\ \vdots \\ X_{gn_g} \end{pmatrix}_{np\times 1} = \begin{pmatrix} 1_{n_1}\otimes I_p & 1_{n_1}\otimes I_p & & \\ 1_{n_2}\otimes I_p & & 1_{n_2}\otimes I_p & \\ \vdots & & & \ddots \\ 1_{n_g}\otimes I_p & & & 1_{n_g}\otimes I_p \end{pmatrix}_{np\times(g+1)p} \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_g \end{pmatrix}_{(g+1)p\times 1} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{g1} \\ \vdots \\ \varepsilon_{gn_g} \end{pmatrix}_{np\times 1},$$

where $n = \sum_{\ell=1}^g n_\ell$ is the total number of observations. Note that

$$X_{\ell r} = \underbrace{\bar{X}_{\cdot\cdot}}_{=\hat{\mu}} + \underbrace{(\bar{X}_{\ell\cdot} - \bar{X}_{\cdot\cdot})}_{=\hat{\tau}_\ell} + \underbrace{(X_{\ell r} - \bar{X}_{\ell\cdot})}_{=\hat{\varepsilon}_{\ell r}},$$

where $\bar{X}_{\ell\cdot} = \frac{1}{n_\ell}\sum_{r=1}^{n_\ell} X_{\ell r}$ and $\bar{X}_{\cdot\cdot} = \frac{1}{n}\sum_{\ell=1}^g \sum_{r=1}^{n_\ell} X_{\ell r}$. Thereby, the *sum of squares and cross products* (SSP) breaks up into

$$\sum_{\ell=1}^g \sum_{r=1}^{n_\ell} (X_{\ell r} - \bar{X}_{\cdot\cdot})(X_{\ell r} - \bar{X}_{\cdot\cdot})' = \sum_{\ell=1}^g n_\ell (\bar{X}_{\ell\cdot} - \bar{X}_{\cdot\cdot})(\bar{X}_{\ell\cdot} - \bar{X}_{\cdot\cdot})' + \sum_{\ell=1}^g \sum_{r=1}^{n_\ell} (X_{\ell r} - \bar{X}_{\ell\cdot})(X_{\ell r} - \bar{X}_{\ell\cdot})'$$

since the cross terms add up to

$$\sum_{\ell=1}^g \underbrace{\sum_{r=1}^{n_\ell} (X_{\ell r} - \bar{X}_{\ell\cdot})}_{=0_p} (\bar{X}_{\ell\cdot} - \bar{X}_{\cdot\cdot})' = 0_{p\times p}.$$

Source	SSP	d.o.f.
Treatment/between-group	$B=\sum_{\ell=1}^{g} n_\ell(\bar{X}_{\ell\cdot}-\bar{X}_{\cdot\cdot})(\bar{X}_{\ell\cdot}-\bar{X}_{\cdot\cdot})'$	$g-1$
Residual/within-group	$W=\sum_{\ell=1}^{g}\sum_{r=1}^{n_\ell}(X_{\ell r}-\bar{X}_{\ell\cdot})(X_{\ell r}-\bar{X}_{\ell\cdot})'$	$n-g$
Total	$B+W=\sum_{\ell=1}^{g}\sum_{r=1}^{n_\ell}(X_{\ell r}-\bar{X}_{\cdot\cdot})(X_{\ell r}-\bar{X}_{\cdot\cdot})'$	$n-1$

Table 1: One-way MANOVA

We summarize the variations in table 1. The *within* SSP can be expressed as $W=\sum_{\ell=1}^{g}(n_\ell-1)S_\ell$, where $S_\ell=\frac{1}{n_\ell-1}\sum_{r=1}^{n_\ell}(X_{\ell r}-\bar{X}_{\ell\cdot})(X_{\ell r}-\bar{X}_{\ell\cdot})'$ is the sample covariance matrix for the ℓ_{th} treatment, and thus the corresponding d.o.f. is $\sum_{\ell=1}^{g}(n_\ell-1)=n-g$.

The hypothesis of no treatment effects, $H_0:\tau_1=\cdots=\tau_g=0_p$, can be tested using **Wilks' Lambda**

$$\Lambda^*=\frac{\det(W)}{\det(B+W)},$$

which is equivalent to $\Lambda=\left(\frac{\det(\hat{\Sigma})}{\det(\hat{\Sigma}_0)}\right)^{n/2}$ in the likelihood-ratio test, where $\hat{\Sigma}=\frac{1}{n}W$ is the unrestricted MLE and $\hat{\Sigma}_0=\frac{1}{n}(B+W)$ is the restricted MLE. Therefore, we reject H_0 if $\Lambda^*<c$ for some c.

dim of variables	no. of groups	Sampling distribution	for	normal data
$p=1$	$g\geq 2$	$\frac{n-g}{g-1}\frac{1-\Lambda^*}{\Lambda^*}$	$\overset{H_0}{\sim}$	$F_{g-1,n-g}$
$p=2$	$g\geq 2$	$\frac{n-g-1}{g-1}\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}$	$\overset{H_0}{\sim}$	$F_{2(g-1),2(n-g-1)}$
$p\geq 1$	$g=2$	$\frac{n-p-1}{p}\frac{1-\Lambda^*}{\Lambda^*}$	$\overset{H_0}{\sim}$	$F_{p,n-p-1}$
$p\geq 1$	$g=3$	$\frac{n-p-2}{p}\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}$	$\overset{H_0}{\sim}$	$F_{2p,2(n-p-2)}$
$n\gg pg$ (large sample)		$-(n-1-\frac{p+g}{2})\log\Lambda^*$	$\xrightarrow[H_0]{d}$	$\chi^2_{(g-1)p}$

Table 2: Distribution of $\Lambda^*=\det(W)/\det(B+W)$

The s-way ANOVA model with $s\in\mathbb{N}$ may be handled in a similar manner. Suppose that there are two factors under consideration. Proceeding by analogy, the *two-way* balanced MANOVA model reads

$$X_{\ell kr}=\mu+\alpha_\ell+\beta_k+\gamma_{\ell k}+\varepsilon_{\ell kr},\quad \ell=1,\cdots,g;\ k=1,\cdots,b;\ r=1,\cdots,n,$$

where $\sum_{\ell=1}^{g}\alpha_\ell=\sum_{k=1}^{b}\beta_k=\sum_{\ell=1}^{g}\gamma_{\ell k}=\sum_{k=1}^{b}\gamma_{\ell k}=0_p$, and $\varepsilon_{\ell kr}\overset{\text{i.i.d.}}{\sim}\mathcal{N}_p(0_p,\Sigma)$. Note that

$$X_{\ell kr}=\underbrace{\bar{X}_{\cdots}}_{=\hat{\mu}}+\underbrace{(\bar{X}_{\ell\cdot\cdot}-\bar{X}_{\cdots})}_{=\hat{\alpha}_\ell}+\underbrace{(\bar{X}_{\cdot k\cdot}-\bar{X}_{\cdots})}_{=\hat{\beta}_k}+\underbrace{(\bar{X}_{\ell k\cdot}-\bar{X}_{\ell\cdot\cdot}-\bar{X}_{\cdot k\cdot}+\bar{X}_{\cdots})}_{=\hat{\gamma}_{\ell k}}+\underbrace{(X_{\ell kr}-\bar{X}_{\ell k\cdot})}_{=\hat{\varepsilon}_{\ell kr}},$$

where a dot is used to denote averaging over the indicated subscript. Accordingly, the *partition of sums of squares and cross products* (SSP) is

$$\begin{aligned}\sum_{\ell=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(X_{\ell kr}-\bar{X}_{\cdots})(X_{\ell kr}-\bar{X}_{\cdots})'=&\sum_{\ell=1}^{g}bn(\bar{X}_{\ell\cdot\cdot}-\bar{X}_{\cdots})(\bar{X}_{\ell\cdot\cdot}-\bar{X}_{\cdots})'\\&+\sum_{k=1}^{b}gn(\bar{X}_{\cdot k\cdot}-\bar{X}_{\cdots})(\bar{X}_{\cdot k\cdot}-\bar{X}_{\cdots})'\\&+\sum_{\ell=1}^{g}\sum_{k=1}^{b}n(\bar{X}_{\ell k\cdot}-\bar{X}_{\ell\cdot\cdot}-\bar{X}_{\cdot k\cdot}+\bar{X}_{\cdots})(\bar{X}_{\ell k\cdot}-\bar{X}_{\ell\cdot\cdot}-\bar{X}_{\cdot k\cdot}+\bar{X}_{\cdots})'\\&+\sum_{\ell=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(X_{\ell kr}-\bar{X}_{\ell k\cdot})(X_{\ell kr}-\bar{X}_{\ell k\cdot})'.\end{aligned}$$

The variations are summarized in table 3.

Source	SSP	d.o.f.
Factor 1	$\mathrm{SSP}_{\texttt{fac1}} = \sum_{\ell=1}^{g} bn(\bar{X}_{\ell\cdot\cdot} - \bar{X}_{\cdots})(\bar{X}_{\ell\cdot\cdot} - \bar{X}_{\cdots})'$	$g-1$
Factor 2	$\mathrm{SSP}_{\texttt{fac2}} = \sum_{k=1}^{b} gn(\bar{X}_{\cdot k\cdot} - \bar{X}_{\cdots})(\bar{X}_{\cdot k\cdot} - \bar{X}_{\cdots})'$	$b-1$
Interaction	$\mathrm{SSP}_{\texttt{int}} = \sum_{\ell=1}^{g}\sum_{k=1}^{b} n(\bar{X}_{\ell k\cdot} - \bar{X}_{\ell\cdot\cdot} - \bar{X}_{\cdot k\cdot} + \bar{X}_{\cdots})(\bar{X}_{\ell k\cdot} - \bar{X}_{\ell\cdot\cdot} - \bar{X}_{\cdot k\cdot} + \bar{X}_{\cdots})'$	$(g-1)(b-1)$
Residual	$\mathrm{SSP}_{\texttt{res}} = \sum_{\ell=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n} (X_{\ell kr} - \bar{X}_{\ell k\cdot})(X_{\ell kr} - \bar{X}_{\ell k\cdot})'$	$gb(n-1)$
Total	$\mathrm{SSP}_{\mathtt{T}} = \sum_{\ell=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n} (X_{\ell kr} - \bar{X}_{\cdots})(X_{\ell kr} - \bar{X}_{\cdots})'$	$gbn-1$

Table 3: Two-way balanced MANOVA

Ordinarily, the test for interaction $\gamma_{\ell k}$ is conducted *before* the tests for main effects α_ℓ, β_k. If interaction effects exist, the factor effects will not have a clear interpretation. An interaction effect means that the

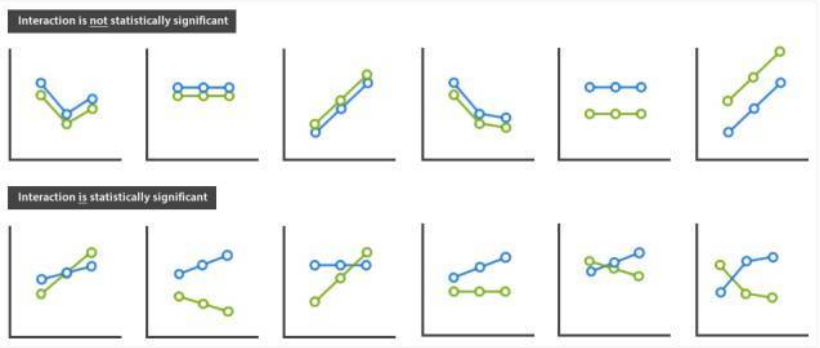

Figure 2: Insignificant interaction (upper half) vs. significant interaction (lower half)

effect of one factor depends on the other factor, and it's indicated by the lines in our profile plot not running *parallel*. Figure 3 shows $4(\texttt{medicine}) \times 2(\texttt{gender}) = 8$ observed mean BDI scores. In this case, the effect

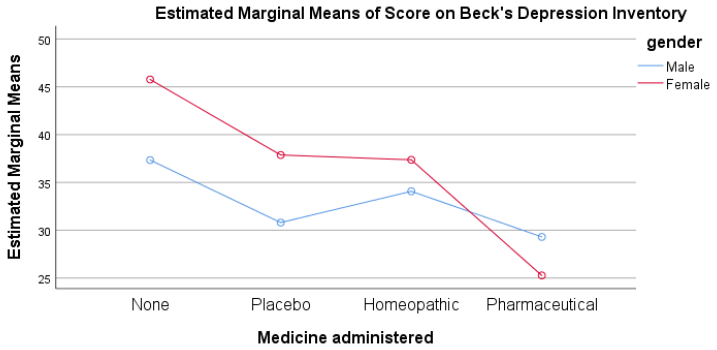

Figure 3: Profile plot of a two-way ANOVA with an interaction effect[xviii)]

for `medicine` interacts with `gender`, i.e., medicine affects females differently from males. Roughly speaking, the red line (Female) descends quite steeply from "None" to "Pharmaceutical" whereas the blue line (Male) is much more horizontal. Depending on `gender`, the main effect of `medicine` obscures –rather than clarifies– how medicine really affects the BDI scores, and thus should be ignored even if it's statistically significant.

[xviii)] from https://www.spss-tutorials.com/spss-two-way-anova-interaction-significant/

First, test $H_0 : \gamma_{\ell k} = 0_p,\ \forall \ell = 1, \cdots, g;\ k = 1, \cdots, b$ using

$$\Lambda^* = \frac{\det(\mathrm{SSP}_{\mathtt{res}})}{\det(\mathrm{SSP}_{\mathtt{int}} + \mathrm{SSP}_{\mathtt{res}})}$$

with

$$-\left[gb(n-1) - \tfrac{p+1-(g-1)(b-1)}{2}\right] \log \Lambda^* \xrightarrow[H_0]{d} \chi^2_{(g-1)(b-1)p}.$$

If $\Lambda^* < c$, reject H_0; otherwise, continue.

Once the full model $X = X_{\mathtt{main}} + X_{\mathtt{int}}$ could be restricted to $X = X_{\mathtt{main}}$, its deserved *orthogonal design*[xix] would allow us to test main effects of Factor 1 and 2 separately.

- $H_{01} : \alpha_1 = \cdots = \alpha_g = 0_p$ may be tested using

$$\Lambda_1^* = \frac{\det(\mathrm{SSP}_{\mathtt{res}})}{\det(\mathrm{SSP}_{\mathtt{fac1}} + \mathrm{SSP}_{\mathtt{res}})}$$

with

$$-\left[gb(n-1) - \tfrac{p+1-(g-1)}{2}\right] \log \Lambda_1^* \xrightarrow[H_0]{d} \chi^2_{(g-1)p}.$$

- $H_{02} : \beta_1 = \cdots = \beta_b = 0_p$ may be tested using

$$\Lambda_2^* = \frac{\det(\mathrm{SSP}_{\mathtt{res}})}{\det(\mathrm{SSP}_{\mathtt{fac2}} + \mathrm{SSP}_{\mathtt{res}})}$$

with

$$-\left[gb(n-1) - \tfrac{p+1-(b-1)}{2}\right] \log \Lambda_2^* \xrightarrow[H_0]{d} \chi^2_{(b-1)p}.$$

§13 Linear Regression (2019/11/14)

To begin with, we discuss the *univariate* regression model

$$y_i = \beta' x_i + \varepsilon_i, \quad i = 1, \cdots, n,$$

where the response y_i is *one-dimensional* and $\beta \in \mathbb{R}^p$ consists of unknown parameters $\beta_1, \ldots, \beta_p$. When the explanatory variables $x_{i1}, \ldots, x_{ip}$ are more than one (i.e., $p > 1$), the model is called *multiple* regression, compared to *simple* linear regression $y_i = a + bz_i + \varepsilon_i$. These n equations are often stacked together and written in matrix notation as

$$y = X\beta + \varepsilon,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Besides, the errors are assumed to satisfy that $\mathbb{E}\varepsilon_i = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for some unknown $\sigma^2 > 0$, and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. That is to say,

$$\mathbb{E}\varepsilon = 0_n \quad \& \quad \mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

The method of least squares yields

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} (y - X\beta)'(y - X\beta) = (X'X)^- X'y.$$

Note that the generalized inverse $(X'X)^-$ of $X'X \in \mathbb{R}^{p \times p}$ may not be well-defined. Nevertheless, the *(ordinary)* **least squares estimator** of β refers unambiguously to

$$\hat{\beta}_{\mathrm{OLS}} = (X'X)^+ X'y = X^+ y,$$

[xix] cf. https://stats.stackexchange.com/q/228797

taking advantage of the unique Moore–Penrose inverse[xx]. Denote $\operatorname{rank}(X) = r \le p$, then $X = X_* C$ for some $X_* \in \mathbb{R}^{n\times r}$ and $C \in \mathbb{R}^{r\times p}$. Let $\beta_* = C\beta \in \operatorname{Col}(C) = \mathbb{R}^r$. It follows that

$$X\beta = X_* C\beta = X_*\beta_*.$$

Note that $\hat\beta_* = (X_*'X_*)^{-1}X_*'y$ is uniquely defined.

In many applications, we are interested in estimating some linear functions $\theta = c'\beta$ of β, where $c \in \mathbb{R}^p$. Recall that ϑ is said to be **estimable** if and only if there exists an *unbiased* estimator of ϑ.

Theorem (estimability in linear models)**.** The following facts suggest that $c'\beta$ is estimable if $c \in \operatorname{Col}(X')$, which is also necessary for $c'\beta$ to be estimable under the assumption that ε is normally distributed.

(1) If $c \in \operatorname{Col}(X')$, then $c'\hat\beta$ is unique and unbiased for $c'\beta$.

(2) If $c'\beta$ is estimable and $\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$, then $c \in \operatorname{Col}(X')$.

Proof. (1) If $c = X'l$ for some $l \in \mathbb{R}^n$, then $c'\hat\beta = l'X(X'X)^-X'y$. It suffices to show

Lemma. The matrix $P_X = X(X'X)^-X'$ is unique. Indeed, P_X is the *orthogonal projection* onto the column space $\operatorname{Col}(X)$ of the design matrix $X \in \mathbb{R}^{n\times p}$.

Proof of Lemma. For any $v \in \mathbb{R}^n$, write $v = x + w$ for $x \in \operatorname{Col}(X)$ and $w \in \operatorname{Col}(X)^\perp$, which exist and are unique since $\mathbb{R}^n = \operatorname{Col}(X) \bigoplus \operatorname{Col}(X)^\perp$. It follows from $X'w = 0_p$ that $P_X w = 0_n$ and $P_X v = P_X x$. To prove $P_X x = x$, it is equivalent to show $\boxed{X(X'X)^-X'X = X}$, or $u'X(X'X)^-X'X = u'X$, $\forall u \in \mathbb{R}^n$. Now that $X'X(X'X)^-X'X = X'X$ by definition, it suffices to point out that $u'X = z'X'X$ for some $z \in \mathbb{R}^p$. What we need is exactly that $\boxed{\operatorname{Col}(X') = \operatorname{Col}(X'X)}$. Clearly $\operatorname{Col}(X'X) \subset \operatorname{Col}(X')$, so the well-known relation $\operatorname{rank}(X'X) = \operatorname{rank}(X) = \operatorname{rank}(X')$ completes the proof. □

Therefore, $c'\hat\beta = l'P_X y$ is unique, and

$$\mathbb{E}[c'\hat\beta] = l'P_X\mathbb{E}y = l'P_X X\beta = l'X\beta = c'\beta.$$

Note that $c = X'l \in \operatorname{Col}(X') \implies \mathbb{E}[l'y] = l'X\beta = c'\beta$.

Remark. We say that $c'\beta$ is *linearly estimable* if, there exists some $l \in \mathbb{R}^n$ such that

$$\mathbb{E}[l'y] = c'\beta,\ \forall\beta \in \mathbb{R}^p.$$

(2) If there is an estimator $T(y, X)$ unbiased for $c'\beta$, then

$$c'\beta = \int_{\mathbb{R}^n} T(y,X)\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right\}\mathrm{d}y.$$

Differentiation with respect to β yields

$$c = X'\int_{\mathbb{R}^n} T(y,X)\frac{y - X\beta}{(2\pi\sigma^2)^{n/2}\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right\}\mathrm{d}y,$$

and thus $c \in \operatorname{Col}(X')$. ■

In summary, $A\beta$ with $A \in \mathbb{R}^{q\times p}$ is estimable iff $\operatorname{Col}(A') \subset \operatorname{Col}(X')$, i.e., $A = A_* X$ for some $A_* \in \mathbb{R}^{q\times n}$. Particularly, β is estimable if and only if X has full column rank.

Sometimes denoted by H as well, the projection matrix $P_X = X(X'X)^-X'$ is also named **hat matrix** as it "puts a hat on y", mapping the vector of response values to the vector of *fitted values*

$$\hat y = X\hat\beta = Hy.$$

The vector of *residuals* is

$$\hat\varepsilon = y - \hat y = (I_n - H)y.$$

Since $HX = X$, $X'H = X'$ and

$$H' = H'H = (H'H)' = H$$

(the check is left to the reader), we have $X'\hat\varepsilon = 0_p$ and $\hat y'\hat\varepsilon = 0$.

[xx] cf. https://zhuanlan.zhihu.com/p/75283604

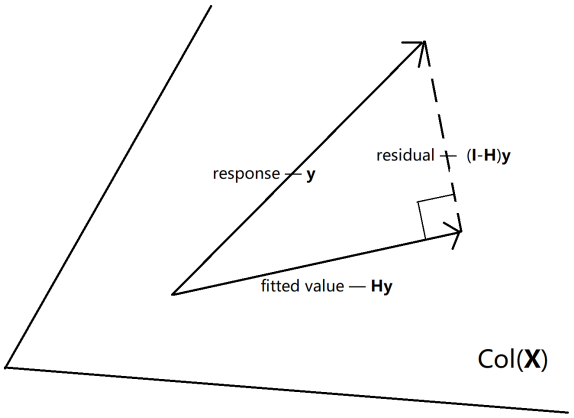

Figure 4: Projection in linear regression

We may view the design matrix from the SVD (see LECTURE §2)

$$\begin{aligned} \underset{(n\times p)}{X} &= \underset{(n\times n)}{U}\,\underset{(n\times p)}{D}\,\underset{(p\times p)}{V'} = \sum_{j=1}^{r} d_j u_j v_j' \\ &= (u_1,\ldots,u_n)\begin{pmatrix}\mathrm{diag}(d_1,\ldots,d_r,0,\cdots,0)\\ 0_{(n-p)\times p}\end{pmatrix}(v_1,\ldots,v_p)'. \end{aligned}$$

The row space of X is $\mathrm{Col}(X') = \mathrm{Col}(X'X) = \mathrm{span}\,(v_1,\ldots,v_r)$, since

$$X'X = \sum_{j=1}^{r} d_j^2 v_j v_j'.$$

The column space of X is $\mathrm{Col}(X) = \mathrm{Col}(XX') = \mathrm{span}\,(u_1,\ldots,u_r)$, since

$$XX' = \sum_{j=1}^{r} d_j^2 u_j u_j'.$$

In addition, $\mathrm{Col}(X)^{\perp} = \mathrm{span}\,(u_{r+1},\ldots,u_n)$. Moreover,

$$P_X = X(X'X)^{-}X' = \sum_{j=1}^{r} u_j u_j' = U_r U_r',$$

and thus

$$P_X y = \sum_{j=1}^{r} \langle y, u_j\rangle\, u_j, \quad \& \quad (I_n - P_X)y = (UU' - U_rU_r')y = \sum_{j=r+1}^{n} \langle y, u_j\rangle\, u_j.$$

The above results are extendable to infinite-dimensional separable Hilbert spaces, e.g., Fourier expansions.

In *multivariate* regression,

$$\underset{(p\times 1)}{Y_i} = \underset{(r\times p)}{\beta}{}' \underset{(r\times 1)}{Z_i} + \underset{(p\times 1)}{U_i}, \quad i = 1,\cdots,n,$$

where $U_i \overset{\text{i.i.d.}}{\sim} (0_p, \Sigma)$ for some unknown $\Sigma = (\sigma_{jk}) \in \mathbb{R}^{p\times p}$. Denote

$$Y = \begin{pmatrix} Y_1' \\ \vdots \\ Y_n' \end{pmatrix} = (Y_{(1)},\ldots,Y_{(p)}), \quad Z = \begin{pmatrix} Z_1' \\ \vdots \\ Z_n' \end{pmatrix}, \quad \beta = \underset{(r\times 1)}{(\beta_{(1)},\ldots,\beta_{(p)})}, \quad U = \begin{pmatrix} U_1' \\ \vdots \\ U_n' \end{pmatrix} = (U_{(1)},\ldots,U_{(p)}).$$

Henceforth,

$$\underset{(n\times p)}{Y} = \underset{(n\times r)}{Z}\underset{(r\times p)}{\beta} + \underset{(n\times p)}{U},$$

where U is a data matrix from $(0_p, \Sigma)$. Rewrite

$$\underset{(n\times 1)}{Y_{(j)}} = \underset{(n\times r)}{Z}\underset{(r\times 1)}{\beta_{(j)}} + \underset{(n\times 1)}{U_{(j)}}, \quad j = 1, \cdots, p,$$

where $\mathrm{Var}(U_{(j)}) = \sigma_{jj} I_n$ and $\mathrm{Cov}(U_{(j)}, U_{(k)}) = \sigma_{jk} I_n$. The method of least squares yields $\hat{\beta}_{(j)} = (Z'Z)^- Z'Y_{(j)}$, so

$$\hat{\beta} = \left(\hat{\beta}_{(1)}, \ldots, \hat{\beta}_{(p)}\right) = (Z'Z)^- Z'Y$$

minimizes the *total variation*

$$\beta \mapsto \mathrm{tr}\left((Y - Z\beta)'(Y - Z\beta)\right) = \sum_{j=1}^{p} (Y_{(j)} - Z\beta_{(j)})'(Y_{(j)} - Z\beta_{(j)})$$

of all observations. The *fitted values* are

$$\hat{Y} = Z\hat{\beta} = Z(Z'Z)^- Z'Y = P_Z Y.$$

The *residuals* are

$$\hat{U} = Y - \hat{Y} = (I_n - P_Z)Y.$$

It's easily seen that

$$Z'\hat{U} = 0_{r\times p}, \text{ i.e., } Z \perp \hat{U}_{(j)};$$

and

$$\hat{Y}'\hat{U} = 0_{p\times p}, \text{ i.e., } \hat{Y}_{(j)} \perp \hat{U}_{(k)}$$

for all $1 \le j, k \le p$. Furthermore, we have the **decomposition of sum of squares and cross products**

$$\underset{\text{(total)}}{Y'Y} = \underset{\text{(regression)}}{\hat{Y}'\hat{Y}} + \underset{\text{(residual)}}{\hat{U}'\hat{U}}.$$

For instance, $\hat{Y} = \frac{1}{n} 1_n 1_n' Y = \bar{Y} 1_n$ if $Z = 1_n$, and all formulae above may come into play.

Assume that the design matrix $Z \in \mathbb{R}^{n\times r}$ has full column rank, in which case $\beta = (\beta_{(1)}, \ldots, \beta_{(p)}) \in \mathbb{R}^{r\times p}$ is estimable. We now investigate properties of $\hat{\beta} = (Z'Z)^{-1}Z'Y$, the least squares estimator. First,

$$\mathbb{E}\hat{\beta} = (Z'Z)^{-1}Z'\mathbb{E}Y = (Z'Z)^{-1}Z'Z\beta = \beta,$$

which demonstrates that $\hat{\beta}$ is *unbiased* for β, and thus $\mathbb{E}\hat{U} = 0_{n\times p}$. Second,

$$\mathrm{Cov}(\hat{\beta}_{(j)}, \hat{\beta}_{(k)}) = (Z'Z)^{-1}Z' \,\mathrm{Cov}(U_{(j)}, U_{(k)}) Z(Z'Z)^{-1} = \sigma_{jk}(Z'Z)^{-1}$$

for $1 \le j, k \le p$. Third,

$$\mathrm{Cov}(\hat{\beta}_{(j)}, \hat{U}_{(k)}) = (Z'Z)^{-1}Z' \,\mathrm{Cov}(U_{(j)}, U_{(k)})(I_n - Z(Z'Z)^{-1}Z') = \sigma_{jk}(Z'Z)^{-1}(Z' - Z') = 0_{r\times n}.$$

When it comes to error terms, note that

$$\mathbb{E}[\hat{U}_{(j)}'\hat{U}_{(k)}] = \mathbb{E}[U_{(j)}'(I_n - P_Z)U_{(k)}] = \mathrm{tr}\left((I_n - P_Z)\mathbb{E}[U_{(k)}U_{(j)}']\right) = \sigma_{kj}\,\mathrm{tr}\,(I_n - P_Z) = (n - \mathrm{rank}\,(Z))\sigma_{jk}.$$

Thus, $\frac{1}{n-\mathrm{rank}(Z)}\mathbb{E}[\hat{U}'\hat{U}] = \Sigma$, and $\hat{\Sigma} = \frac{1}{n-\mathrm{rank}(Z)}\hat{U}'\hat{U}$ is called the *within-sample MSE*[xxi] *of the predictor*. Prediction of a new observation is more uncertain than estimating the expected value. See also exercise 7 in homework 1. Consider

$$\underset{\text{(new response)}}{Y_0} = \underset{\text{(expected value)}}{\beta' z_0} + \underset{\text{(new error)}}{U_0}$$

at $z_0 \in \mathbb{R}^r$. Suppose that $Z \in \mathbb{R}^{n\times r}$ has full column rank, then $z_0'\hat{\beta}$ is an unbiased predictor of $z_0'\beta$. The *forecast error* is

$$Y_0 - \hat{\beta}' z_0 = (\beta - \hat{\beta})' z_0 + U_0,$$

whose covariance matrix consists of

$$\mathrm{Cov}(z_0'(\beta_{(j)} - \hat{\beta}_{(j)}), z_0'(\beta_{(k)} - \hat{\beta}_{(k)})) + \mathrm{Cov}(U_{0,j}, U_{0,k}) = \sigma_{jk}[1 + z_0'(Z'Z)^{-1}z_0], \quad 1 \le j, k \le p.$$

In short, $\mathrm{Var}(Y_0 - \hat{\beta}' z_0) = [1 + z_0'(Z'Z)^{-1}z_0]\Sigma$, where $z_0'(Z'Z)^{-1}z_0 \ge 0$.

[xxi] the definition of the **mean squared error** differs according to whether one is describing a *predictor* or an *estimator* (see LECTURE §3 for the latter).

§14 Multivariate Linear Regression Cont'd (2019/11/19)

From now on, we assume, WLOG, that Z, the design matrix in the linear model

$$\underset{(n\times p)}{Y} = \underset{(n\times r)}{Z}\underset{(r\times p)}{\beta} + \underset{(n\times p)}{U},$$

has full column rank so that β is estimable (see LECTURE §13); otherwise, we may pick out some maximal linearly independent columns of Z, resulting in *dimensionality reduction* of covariates. Denote by

$$\hat{\beta} = (Z'Z)^{-1}Z'Y = \beta + (Z'Z)^{-1}Z'U$$

the least squares estimator (**LSE**) of β. Note that the rows of the error term U are uncorrelated zero-mean random vectors, whose covariance matrices are identically $\Sigma \in \mathbb{R}^{p\times p}$, an unknown *nuisance parameter*.

The **Gauss–Markov theorem** claims that $\hat{\beta}$ is the *best linear unbiased estimator* (BLUE) of β, in the sense that for any $c'\beta$ ($c \in \mathbb{R}^r$), of the estimators $l'Y$ ($l \in \mathbb{R}^n$) that are unbiased, $c'\hat{\beta}$ has minimum covariance matrix in Loewner order.

Proof. Since $c'\beta = \mathbb{E}[l'Y] = l'Z\beta$ for all β, we have $c = Z'l$. To show that $\operatorname{vec}(c'\hat{\beta}) = \operatorname{vec}(l'P_ZY) = (I_p \otimes (l'P_Z))\operatorname{vec}(Y)$ has smaller covariance matrix than $\operatorname{vec}(l'Y) = (I_p \otimes l')\operatorname{vec}(Y)$, direct calculations yield

$$\operatorname{Var}(\operatorname{vec}(c'\hat{\beta})) = (I_p \otimes (l'P_Z))(\Sigma \otimes I_n)(I_p \otimes (l'P_Z))' = (l'P_Zl)\Sigma$$

and $\operatorname{Var}(\operatorname{vec}(l'Y)) = (l'l)\Sigma$. It suffices to observe that $l'(I_n - P_Z)l \geq 0$. □

Hereafter, assume further that the rows of U are i.i.d. random vectors drawn from $\mathcal{N}_p(0_p, \Sigma)$, i.e., $\operatorname{vec}(U) \sim \mathcal{N}_{np}(0_{np}, \Sigma \otimes I_n)$. In such a Gaussian framework, the MLE of β is identical to its LSE. Actually, the likelihood function

$$L_n(\beta, \Sigma | Y_1, \ldots, Y_n) = (2\pi)^{-np/2}\det(\Sigma)^{-n/2}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\sum_{i=1}^n (Y_i - \beta'Z_i)(Y_i - \beta'Z_i)'\right)\right\}$$

attains its maximum at

$$(\hat{\beta}^{\mathrm{MLE}}, \hat{\Sigma}^{\mathrm{MLE}}) = (\hat{\beta}, \tfrac{1}{n}\hat{U}'\hat{U}),$$

where $\hat{U} = Y - \hat{Y} = (I_n - P_Z)Y$ has $\hat{U}_i = Y_i - \hat{\beta}'Z_i$ as its rows. To see this, notice that the SSP of

$$Y_i - \beta'Z_i = \hat{U}_i + (\hat{\beta} - \beta)'Z_i$$

is

$$\sum_{i=1}^n (Y_i - \beta'Z_i)(Y_i - \beta'Z_i)' = \sum_{i=1}^n \hat{U}_i\hat{U}_i' + \sum_{i=1}^n (\hat{\beta} - \beta)'Z_iZ_i'(\hat{\beta} - \beta) = \hat{U}'\hat{U} + (\hat{\beta} - \beta)'Z'Z(\hat{\beta} - \beta),$$

since the cross terms add up to

$$\sum_{i=1}^n \hat{U}_iZ_i'(\hat{\beta} - \beta) = \hat{U}'Z(\hat{\beta} - \beta) = 0_{p\times r}(\hat{\beta} - \beta) = 0_{p\times p}.$$

Combined with exercise 2 in homework 2, the decomposition of SSP makes the expressions of $\hat{\beta}^{\mathrm{MLE}}$ and $\hat{\Sigma}^{\mathrm{MLE}}$ clear. Since $\hat{\Sigma}^{\mathrm{MLE}}$ is biased, the unbiased $\hat{\Sigma}^{\mathrm{MoM}} = \frac{1}{n-r}\hat{U}'\hat{U}$ (*method of moments* estimator) prevails in linear regression. Note that

$$n\hat{\Sigma}^{\mathrm{MLE}} = Y'(I_n - Z(Z'Z)^{-1}Z')Y = (Y - Z\beta)'(I_n - P_Z)(Y - Z\beta) = U'(I_n - P_Z)U \sim W_p(\Sigma, n - r),$$

not depending on β.

After estimation, let's get down to inferences, say, the likelihood-ratio test (LRT). Write $\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix}$ and $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ such that $\boldsymbol{\beta}_1 \in \mathbb{R}^{q\times p}$, $\boldsymbol{\beta}_2 \in \mathbb{R}^{(r-q)\times p}$, $\mathbf{Z}_1 \in \mathbb{R}^{n\times q}$, and $\mathbf{Z}_2 \in \mathbb{R}^{n\times(r-q)}$ for some $q < r$. Then,

$$\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\beta}_1 + \mathbf{Z}_2\boldsymbol{\beta}_2 + \mathbf{U}.$$

We are interested in $H_0 : \boldsymbol{\beta}_2 = 0_{(r-q)\times p}$. Since $\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\beta}_1 + \mathbf{U}$ under H_0, similar arguments give constrained MLEs

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Y}$$

and

$$\hat{\Sigma}_1^{\text{MLE}} = \tfrac{1}{n}(\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1)'(\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1) = \tfrac{1}{n}\mathbf{Y}'(I_n - P_{\mathbf{Z}_1})\mathbf{Y}.$$

The LRT statistic is

$$\Lambda = \frac{L_n(\hat{\boldsymbol{\beta}}_1, \hat{\Sigma}_1^{\text{MLE}})}{L_n(\hat{\boldsymbol{\beta}}, \hat{\Sigma}^{\text{MLE}})} = \frac{(2\pi)^{-np/2}\det(\hat{\Sigma}_1^{\text{MLE}})^{-n/2}\exp(-\frac{np}{2})}{(2\pi)^{-np/2}\det(\hat{\Sigma}^{\text{MLE}})^{-n/2}\exp(-\frac{np}{2})} = \left(\frac{\det(\hat{\Sigma}^{\text{MLE}})}{\det(\hat{\Sigma}_1^{\text{MLE}})}\right)^{n/2}.$$

Also, Wilks' Lambda

$$\Lambda^* = \Lambda^{2/n} = \frac{\det(\hat{\Sigma}^{\text{MLE}})}{\det(\hat{\Sigma}_1^{\text{MLE}})}$$

is commonly used. Here, the SSP decomposition is

$$\underbrace{(\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1)'(\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1)}_{=n\hat{\Sigma}_1^{\text{MLE}}} = \underbrace{(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}})}_{=n\hat{\Sigma}^{\text{MLE}}} + \underbrace{(\mathbf{Z}\hat{\boldsymbol{\beta}} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1)'(\mathbf{Z}\hat{\boldsymbol{\beta}} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1)}_{=n(\hat{\Sigma}_1^{\text{MLE}} - \hat{\Sigma}^{\text{MLE}})},$$

where the cross terms add up to

$$(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{Z}\hat{\boldsymbol{\beta}} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_1) = \mathbf{Y}'(I_n - P_{\mathbf{Z}})(P_{\mathbf{Z}} - P_{\mathbf{Z}_1})\mathbf{Y} = 0_{p\times p},$$

since $\text{Col}(\mathbf{Z}_1) \subset \text{Col}(\mathbf{Z})$. From the perspective of Cochran's theorem (see LECTURE §9),

$$I_n = P_{\mathbf{Z}_1} + (P_{\mathbf{Z}} - P_{\mathbf{Z}_1}) + (I_n - P_{\mathbf{Z}})$$

decomposes into mutually orthogonal projections onto $\text{Col}(\mathbf{Z}_1)$, $\text{Col}(\mathbf{Z}) \cap \text{Col}(\mathbf{Z}_1)^\perp$, and $\text{Col}(\mathbf{Z})^\perp$, respectively. Therefore,

$$n\hat{\Sigma}_1^{\text{MLE}} \overset{H_0}{\sim} W_p(\Sigma, n-q), \qquad \& \qquad n(\hat{\Sigma}_1^{\text{MLE}} - \hat{\Sigma}^{\text{MLE}}) \overset{H_0}{\sim} W_p(\Sigma, r-q).$$

Moreover, the large sample approximation is

$$-2\log\Lambda \approx -\left[n - r - 1 - \tfrac{1}{2}(p - r + q + 1)\right]\log\left(\frac{\det(\hat{\Sigma}^{\text{MLE}})}{\det(\hat{\Sigma}_1^{\text{MLE}})}\right) \overset{\circ}{\sim} \chi^2_{(r-q)p}.$$

Generally, to test $H_0 : \underset{(r-q)\times r}{C}\beta = \underset{(r-q)\times p}{A}$, we can use the fact that

$$n(\hat{\Sigma}_0^{\text{MLE}} - \hat{\Sigma}^{\text{MLE}}) = (C\hat{\beta} - A)'[C(Z'Z)^{-1}C']^{-1}(C\hat{\beta} - A) \overset{H_0}{\sim} W_p(\Sigma, r-q),$$

and $n\hat{\Sigma}_0^{\text{MLE}} \overset{H_0}{\sim} W_p(\Sigma, n-q)$.

Recall that

$$T^2 = d'(\frac{M}{m})^{-1}d \sim T^2(p,m) = \frac{mp}{m-p+1}F_{p,m-p+1}$$

when $d \sim \mathcal{N}_p(0_p, \Sigma) \perp\!\!\!\perp M \sim W_p(\Sigma, m)$. For any $z_0 \in \mathbb{R}^r$, we have

$$\hat{\beta}'z_0 \sim \mathcal{N}_p(\beta'z_0, z_0'(Z'Z)^{-1}z_0\Sigma) \quad \perp\!\!\!\perp \quad n\hat{\Sigma}^{\text{MLE}} \sim W_p(\Sigma, n-r),$$

and thus

$$\left(\frac{\hat{\beta}'z_0 - \beta'z_0}{\sqrt{z_0'(Z'Z)^{-1}z_0}}\right)'\left(\frac{n\hat{\Sigma}^{\text{MLE}}}{n-r}\right)^{-1}\left(\frac{\hat{\beta}'z_0 - \beta'z_0}{\sqrt{z_0'(Z'Z)^{-1}z_0}}\right) \sim T^2(p, n-r) = \frac{(n-r)p}{n-r-p+1}F_{p,n-r-p+1}.$$

In the course of predicting/forecasting $Y_0 = \beta'z_0 + U_0$, we have

$$Y_0 - \hat{\beta}'z_0 = (\beta - \hat{\beta})'z_0 + U_0 \sim \mathcal{N}_p(0_p, [1 + z_0'(Z'Z)^{-1}z_0]\Sigma) \quad \perp\!\!\!\perp \quad n\hat{\Sigma}^{\text{MLE}} \sim W_p(\Sigma, n-r),$$

and thus

$$(Y_0 - \hat{\beta}'z_0)'\left(\frac{n\hat{\Sigma}^{\text{MLE}}}{n-r}\right)^{-1}(Y_0 - \hat{\beta}'z_0) \sim [1 + z_0'(Z'Z)^{-1}z_0]\frac{(n-r)p}{n-r-p+1}F_{p,n-r-p+1}.$$

§15 Model Diagnostics and Selection (2019/11/26)

Recall the univariate linear model (see LECTURE §13)

$$\underset{(n\times 1)}{y} = \underset{(n\times p)}{X}\underset{(p\times 1)}{\beta} + \underset{(n\times 1)}{\varepsilon}.$$

The vector of residuals is

$$\hat{\varepsilon} = y - \hat{y} = (I_n - X(X'X)^{-}X')y.$$

Denote

$$H = X(X'X)^{-}X' = (h_{ij})_{1\le i,j\le n}.$$

Since $H = H' = H^2$, we have

$$h_{ij} = \sum_{k=1}^{n} h_{ik}h_{kj} = \sum_{k=1}^{n} h_{ik}h_{jk}.$$

The diagonal elements of H are called the **leverages** (i.e., influential points). Note that

$$\sum_{i=1}^{n} h_{ii} = \operatorname{tr}(H) = \operatorname{rank}(H) = \operatorname{rank}(X),$$

and

$$h_{ii}(1-h_{ii}) = \sum_{j\neq i} h_{ij}^2 \ge 0 \implies h_{ii} \in [0,1].$$

If $h_{ii} \nearrow 1$, then

$$\hat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$$

reflects more contribution from y_i. However, for prediction, we hope that y_j $(j \neq i)$ contribute to $\hat{y}_i$ as much as possible, while y_i contributes to $\hat{y}_i$ as even as possible.

It is imperative to examine the *adequacy* of the model before making decisions. Since

$$\operatorname{Var}(\hat{\varepsilon}) = (I_n - H)\operatorname{Var}(y)(I_n - H)' = (I_n - H)(\sigma^2 I_n)(I_n - H) = \sigma^2(I_n - H),$$

we often prefer **studentized residuals**

$$\hat{\varepsilon}_i^* = \frac{\hat{\varepsilon}_i}{\sqrt{s^2(1-h_{ii})}}, \quad i = 1, \cdots, n$$

using the mean squared residual $s^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ as an estimate of σ^2. Also, we may studentize $\hat{\varepsilon}_i$ using the delete-one estimated variance $s_{(-i)}^2$, computed from the regression with the i^{th} observation dropped. The following **graphical diagnostics** are based on these studentized residuals:

- Q-Q plots and histograms for normality.
- Scatter plots of $\hat{\varepsilon}_i^*$ vs. i (indexes) for temporal/spatial correlation.
- Scatter plots of $\hat{\varepsilon}_i^*$ vs. $\hat{y}_i$ for homoscedasticity.
- Scatter plots of $\hat{\varepsilon}_i^*$ vs. X_i for flexible fit (e.g., need for more terms in the model).

Note that there exist some subjective judgements. A good sense comes from practices!

In spite of most realistic situations that the true model is not in our candidate class, we shall try our best to make our model useful, i.e., to approximate the true model, which often involves nonparametric (infinite-dimensional) methods. If, coincidentally, the true model belongs to the candidate class, then our goal is simply to find it out, in which case we often utilize parametric (finite-dimensional) methods and pay attention to consistency. These are the so called **model selection**. For example, we often use penalties or shrinkage in high-dimensional statistics ($p \gg n$), where models rest on a few *tuning parameters*. For the purpose of selection, some criteria are needed. Note that models, offering ways of inference for the population feature, ultimately serve the prediction.

Suppose the true model is

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\mu} \in \mathbb{R}^n, \ \boldsymbol{\varepsilon} \sim (0_n, \sigma^2 I_n).$$

The candidate class consists of stepwise regression models

$$\mathbf{y} = \mathbf{X}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon}, \quad p = 1, \cdots, m$$

where $\mathbf{X}_p \in \mathbb{R}^{n\times p}$ form a nested sequence $\mathbf{X}_1 \subset \mathbf{X}_2 \subset \cdots \subset \mathbf{X}_m$, e.g., splines with more and more knots, or Fourier expansions of higher and higher order. The method of least squares yields $\hat{\boldsymbol{\beta}}_p = (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{y}$, and thereby

$$\hat{\boldsymbol{\mu}}_p = \mathbf{X}_p\hat{\boldsymbol{\beta}}_p = \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{y} = \mathbf{P}_p\mathbf{y}.$$

Clearly, the bias

$$\mathbf{b}_p = \boldsymbol{\mu} - \mathbb{E}\hat{\boldsymbol{\mu}}_p = (I_n - \mathbf{P}_p)\boldsymbol{\mu}$$

decreases in p. Particularly, if $\boldsymbol{\mu} = \mathbf{X}_{p^*}\boldsymbol{\beta}_{p^*}$ for some p^*, i.e., the true model belongs to the candidate class, then $\mathbf{b}_p = 0_n$ for all $p \geq p^*$. Besides, the mean squared error of $\hat{\boldsymbol{\mu}}_p$,

$$d_p := \mathbb{E}\left\|\hat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}\right\|^2 = \mathbb{E}\left\|\mathbf{P}_p\boldsymbol{\varepsilon} - (I_n - \mathbf{P}_p)\boldsymbol{\mu}\right\|^2 = \mathbb{E}\left\|\mathbf{P}_p\boldsymbol{\varepsilon}\right\|^2 + \left\|(I_n - \mathbf{P}_p)\boldsymbol{\mu}\right\|^2 = p\sigma^2 + \left\|\mathbf{b}_p\right\|^2,$$

decomposes into its variance $p\sigma^2$ and its squared bias $\left\|\mathbf{b}_p\right\|^2$. Generally, the optimal p can be estimated by

$$\hat{p} = \underset{1\leq p\leq m}{\arg\min}\, \hat{d}_p,$$

where the estimate $\hat{d}_p$ of d_p should be specified. Since the sum of squared errors for the candidate model with p covariates,

$$\mathrm{SSE}_p = \left\|\mathbf{y} - \mathbf{X}_p\hat{\boldsymbol{\beta}}_p\right\|^2 = \left\|(I_n - \mathbf{P}_p)(\boldsymbol{\mu} + \boldsymbol{\varepsilon})\right\|^2 = \left\|\mathbf{b}_p + (I_n - \mathbf{P}_p)\boldsymbol{\varepsilon}\right\|^2,$$

satisfies that

$$\mathbb{E}\mathrm{SSE}_p = \mathbb{E}\left\|(I_n - \mathbf{P}_p)\boldsymbol{\varepsilon}\right\|^2 + \left\|\mathbf{b}_p\right\|^2 = (n-p)\sigma^2 + \left\|\mathbf{b}_p\right\|^2,$$

we can define

$$\hat{d}_p = \mathrm{SSE}_p - (n-2p)\hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is unbiased so that $\mathbb{E}\hat{d}_p = d_p$. A similar criterion is Akaike's **final prediction error** (FPE). Imagine a vector of new data $\mathbf{y}^* = \boldsymbol{\mu} + \boldsymbol{\varepsilon}^*$. The expectation of the sum of squared prediction errors is

$$d_p := \mathbb{E}\mathrm{PE}_p = \mathbb{E}\left\|\mathbf{y}^* - \hat{\boldsymbol{\mu}}_p\right\|^2 = n\sigma^2 + \mathbb{E}\left\|\hat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}\right\|^2 = (n+p)\sigma^2 + \left\|\mathbf{b}_p\right\|^2.$$

Hence, we may take

$$\hat{d}_p = \mathrm{FPE}_p = \mathrm{SSE}_p + 2p\hat{\sigma}^2,$$

and then we are faced with the trade-off between SSE_p and $2p\sigma^2$. So, how to estimate σ^2?

1) If we choose $\hat{\sigma}^2 = \mathrm{MSE}_p = \mathrm{SSE}_p/(n-p)$, then

$$\hat{d}_p = \mathrm{SSE}_p + 2p\mathrm{MSE}_p = \tfrac{n+p}{n-p}\mathrm{SSE}_p.$$

For underfit models, the MSE_p can be dramatically larger than σ^2. Thus, a large $\hat{p}$ will be encouraged.

2) If we choose $\hat{\sigma}^2 = \mathrm{MSE}_m$ typically, using the largest model, then

$$\hat{d}_p = \mathrm{SSE}_p + 2p\tfrac{\mathrm{SSE}_m}{n-m},$$

which, however, cannot capture consistent models.

One of the most popular criterion,

$$C_p = \tfrac{1}{n}(\mathrm{SSE}_p + 2p\hat{\sigma}^2),$$

is called **Mallow's** C_p (with $\hat{\sigma}^2 = \mathrm{MSE}_m$), sometimes defined as

$$C_p' = \mathrm{SSE}_p/\hat{\sigma}^2 - (n-2p) = nC_p/\hat{\sigma}^2 - n.$$

Other criteria to be minimized among $1 \leq p \leq m$ include

- **Akaike information criterion**
$$\text{AIC} = -2\log(\hat{L}_n) + 2p,$$
where $\hat{L}_n$ is the maximum value of the likelihood function of the model with p covariates. In the case of (Gaussian) linear regression, ignoring the constant terms allows us to conveniently take
$$\text{AIC} = n\log(\text{SSE}_p/n) + 2p,$$
which is asymptotically equivalent to Mallow's C_p.

- **Bayesian information criterion** (developed by Schwarz who gave a Bayesian argument)
$$\text{BIC} = -2\log(\hat{L}_n) + \log(n)p,$$
where $\hat{L}_n$ is the maximum value of the likelihood function of the model with p covariates. In the case of (Gaussian) linear regression, ignoring the constant terms allows us to conveniently take
$$\text{BIC} = n\log(\text{SSE}_p/n) + \log(n)p.$$

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

§16 Model Selection Cont'd & Principal Components (2019/11/28)

Let $\hat{p}_M$ be the model chosen by Mallow's C_p, which is asymptotically equivalent to AIC, and let $\hat{p}_S$ be the model chosen by Schwarz's BIC. That is,
$$\hat{p}_M = \underset{p\in\{1,\cdots,m\}}{\arg\min}\ C_p, \quad \text{and} \quad \hat{p}_S = \underset{p\in\{1,\cdots,m\}}{\arg\min}\ \text{BIC}_p.$$
Their optimalities are as follows. Assume that $m \to \infty$ as $n \to \infty$.

- If $\boldsymbol{\mu} \neq \mathbf{X}_p\boldsymbol{\beta}_p$ for $p = 1, \cdots, m$, i.e., the true model is not in the candidate class, then
$$\frac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\hat{p}_M}\|}{\inf_{1\le p\le m}\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_p\|} \xrightarrow{\mathbb{P}} 1, \quad \text{and} \quad \frac{d_{\hat{p}_M}}{\inf_{1\le p\le m} d_p} \xrightarrow{\mathbb{P}} 1.$$

- If $\boldsymbol{\mu} = \mathbf{X}_{p^*}\boldsymbol{\beta}_{p^*}$ for some p^*, i.e., the true model belongs to the candidate class, then
$$\mathbb{P}\{\hat{p}_S = p^*\} \to 1, \quad \text{while} \quad \mathbb{P}\{\hat{p}_M > \hat{p}_S\} > 0.$$

A widely used method for assessing the performance of prediction models is **cross-validation** (CV). In K-fold cross-validation, the original sample $(x_i, y_i)_{1\le i\le n}$ is randomly partitioned into K equal-sized subsamples. Let $\kappa : \{1, \cdots, n\} \to \{1, \cdots, K\}$ be an indexing function that indicates the partition to which observations are allocated by the randomization, respectively. Denote by $\hat{f}_{-k}(\cdot)$ the fitted function, computed with the k^{th} part of the data removed. Then the *estimate of prediction error* is
$$\text{CV} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n/K}\sum_{\kappa(i)=k}\left[y_i - \hat{f}_{-k}(x_i)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_{-\kappa(i)}(x_i)\right]^2.$$
In practice, one typically performs K-fold CV using $K = 5$ or $K = 10$, which is computationally efficient. The case $K = n$ is known as *leave-one-out* cross-validation, where
$$\text{LOOCV} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_{(-i)}(x_i)\right]^2.$$
Consider the linear regression model with p regressors, where $f(x) = x'\beta$ and $y - f(x) = \varepsilon \sim (0, \sigma^2)$. Clearly we have $\hat{f}(x) = x'\hat{\beta}$ and $\hat{f}_{(-i)}(x) = x'\hat{\beta}_{(-i)}$ by the method of least squares. Forecasting a new response y^* at x^* gives prediction error $\text{PE} = \left[y^* - \hat{f}(x^*)\right]^2$. Then it holds that $\mathbb{E}\text{LOOCV} = \mathbb{E}\text{PE} + O(p/n^2)$.

Generally, LOOCV can be easily calculated for ridge-type regression, where the hat matrix takes the form

$$H = X(X'X + \lambda J)^{-1}X = (h_{ij})_{1\le i,j\le n}.$$

The goal is

$$\text{LOOCV} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(x_i)}{1-h_{ii}}\right]^2,$$

so it suffices to show that

$$y_i - \hat{f}_{(-i)}(x_i) = \frac{y_i - \hat{f}(x_i)}{1-h_{ii}}.$$

Proof. Let $\tilde{y} = (\tilde{y}_j)_{1\le j\le n} = \left(y_1, \ldots, y_{i-1}, \hat{f}_{(-i)}(x_i), y_{i+1}, \ldots, y_n\right)'$. Then the minimizer $\hat{f}_{(-i)}(\cdot)$ of the leave-one-out penalized sum of squares $\sum_{j\neq i}[y_j - f(x_j)]^2 + \lambda\mathcal{J}(f)$ also minimizes $\sum_{j=1}^{n}[\tilde{y}_j - f(x_j)]^2 + \lambda\mathcal{J}(f)$. It follows that $\left(\hat{f}_{(-i)}(x_j)\right)_{1\le j\le n} = H\tilde{y}$. Just combine $\hat{f}_{(-i)}(x_i) = \sum_{j=1}^{n} h_{ij}\tilde{y}_j$ with $\hat{f}(x_i) = \sum_{j=1}^{n} h_{ij}y_j$. □

It's immediate that LOOCV is not stable if some $h_{ii} \nearrow 1$. To alleviate the tendency to undersmooth, a feasible approximation is **generalized cross-validation**, given as

$$\text{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(x_i)}{1-\operatorname{tr}(H)/n}\right]^2.$$

Using the fact that $(1-p/n)^{-2} \approx (n+p)/(n-p)$ when $\operatorname{tr}(H) = p \ll n$, we have

$$\text{GCV} = \frac{1}{n}\left(1-\frac{p}{n}\right)^{-2}\sum_{i=1}^{n}\left[y_i - \hat{f}(x_i)\right]^2 \approx \frac{1}{n}\frac{n+p}{n-p}\text{SSE} = \frac{1}{n}\text{FPE}.$$

The above discussions can be extended to multivariate cases, e.g., $\begin{Bmatrix}\text{AIC}\\ \text{BIC}\end{Bmatrix} = n\log(\det\hat{\Sigma}_d^{\text{MLE}}) + \begin{Bmatrix}2\\ \log(n)\end{Bmatrix} pd$ for Gaussian linear regression of p-dimensional responses with d covariates.

For computational and statistical reasons, *best subset* selection may suffer from an enormous search space. Thus, *stepwise* methods are attractive alternatives. Hybrid versions of forward and backward stepwise selection attempts to more closely mimic best subset selection while retaining the computational advantages of forward and backward stepwise selection. Rather than subset selection methods, we may fit a model containing all predictors using a technique that *regularizes* the coefficient estimates, i.e., we may introduce a shrinkage penalty and select the tuning parameter.

In multivariate statistics, *data reduction* plays a central role, which transforms the predictors and reproduces the variability. **Principal components analysis** (PCA) is commonly used for deriving a low-dimensional set of features from a large set of variables. As for intuition, $(x_i, 0)_{1\le i\le n}$ should be extracted from its perturbed realization $(x_i, u_i)_{1\le i\le n}$, where $|u_i| \ll 1$. For any data cloud, we should find some directions onto which the projection of the data shows the most information/variation.

From the point of view of population, let X be a random p-vector and suppose that $\mathbb{E}X = 0_p$ for simplicity. We are to find some unit vector $\ell_1 \in \mathbb{R}^p$ such that $Y_1 = \ell_1'X$ captures most of the variability. Indeed,

$$\ell_1 = \underset{\ell\in\mathbb{R}^p:\|\ell\|=1}{\arg\max}\ \operatorname{Var}(\ell'X) = \underset{\ell\in\mathbb{R}^p:\|\ell\|=1}{\arg\max}\ \ell'\Sigma\ell$$

turns out to be the first eigenvector of $\operatorname{Var}(X) = \Sigma$. Sequentially, the j^{th} **principal component** of X is $Y_j = \ell_j'X$, where $\ell_j = \arg\max_{\ell\in\mathbb{R}^p}\ell'\Sigma\ell$ s.t. $\|\ell\| = 1$ and $\ell'\ell_k = 0$ for $1 \le k < j$. Using the eigen-decomposition

$$\Sigma = (e_1, \ldots, e_p)\operatorname{diag}(\lambda_1, \ldots, \lambda_p)(e_1, \ldots, e_p)' = \sum_{j=1}^{p}\lambda_j e_j e_j',$$

where $\lambda_1 > \cdots > \lambda_p$[xxii], we have[xxiii] $Y_j = e_j'X$ and thus $\operatorname{Var}((Y_j)_{1\le j\le p}) = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$.

[xxii] the case when $\lambda_j = \lambda_k$ for some $j \neq k$ (multiplicity) is usually not considered in statistical analysis, since observations are assumed to be drawn randomly.

[xxiii] cf. https://zhuanlan.zhihu.com/p/75433434

§17 Principal Component Analysis Cont'd (2019/12/3)

For unknown $\Sigma = \mathrm{Var}(X)$, we can estimate it by the sample covariance matrix S or the MLE $\hat{\Sigma}$. If $n \gg p$, then the sample principal components (PCs) are consistent in the sense that

$$\{\hat{e}_j, \hat{\lambda}_j\} \to_{\mathbb{P}} \{e_j, \lambda_j\}.$$

One may consider the *random effect model* (linear mixed-effects model)

$$Y_i = \underbrace{X_i'\beta}_{\text{mean (fixed)}} + \underbrace{Z_i'\gamma_i}_{\text{subject-specific (random)}} + \varepsilon_i,$$

where the PCs seem inclined to approximate these fixed effects.

Back to the population view: let

$$Q = (e_1, \ldots, e_p),$$

then

$$Y = (Y_1, \ldots, Y_p)' = Q'X,$$

and

$$X = (X_1, \ldots, X_p)' = QY = \sum_{j=1}^{p} Y_j e_j = \sum_{j=1}^{p} \langle X, e_j \rangle e_j.$$

Generally, for a (non-centered) random element X taking values in some separable Hilbert space with $\mathbb{E}X = \mu$, we have the so called *Karhunen-Loève expansion*[xxiv)]

$$X = \mu + \sum_{j=1}^{\infty} \langle X - \mu, e_j \rangle e_j.$$

Note that the **total variance** is invariant in that

$$\sum_{j=1}^{p} \mathrm{Var}(Y_j) = \sum_{j=1}^{p} \lambda_j = \mathrm{tr}\,(\Sigma) = \sum_{j=1}^{p} \mathrm{Var}(X_j).$$

The unexplained variance of the p-dimensional $X = \mu + \sum_{j=1}^{p} Y_j e_j$, with respect to the p_1-dimensional $X_* = \mu + \sum_{j=1}^{p_1} Y_j e_j$, is

$$\mathbb{E}\|X - X_*\|^2 = \mathbb{E}\left\| \sum_{j=p_1+1}^{p} Y_j e_j \right\|^2 = \sum_{j=p_1+1}^{p} \lambda_j.$$

The **fraction of variation explained** by the first p_1 PCs is then defined to be

$$\mathrm{FVE} = \sum_{j=1}^{p_1} \lambda_j \Big/ \sum_{j=1}^{p} \lambda_j,$$

which reflects the model fidelity. However, p_1 characterizes the model complexity. We again have a trade-off between two goals, as we already encountered with MSEs. The optimal p_1^* is usually chosen with reference to a plot of cumulative FVE against p_1 — there are no universal rules, only rules of thumb. In view of this, the distribution of λ's is essential, whose properties have been intensively studied in random matrix theory. As a result, one often see that $p_1^* \sim n^\alpha$ with a small $\alpha \in (0, 1)$.

PCA can also be done using the correlation matrix instead of the covariance matrix. Note that the correlation matrix is exactly the covariance matrix of the scaled/standardized variables. Results of scaled and unscaled PCA can differ a lot, since scaling makes variables equally important. Often (but not always), scaling (i.e., using the correlation matrix) is preferred.

Anyway, PCA is a technique for dimensionality reduction that pursues a simplified structure of the initial data, and belongs to unsupervised learning. Sometimes the PCs are used in supervised learning, e.g., regression/classification, where the explanation and selection will become much more clear.

The interpretation can be difficult. When all measurements are positively correlated, the first PC is often some kind of *average* of the measurements (e.g., market general index). Then the other PCs give important information about the remaining pattern (e.g., industrial sector index).

[xxiv)] cf. https://zhuanlan.zhihu.com/p/77418542

The geometrical interpretation of PCs based on data is the minimization of perpendicular distances, as shown in Figure 5, not to be confused with vertical distances in OLS, where responses and predictors are

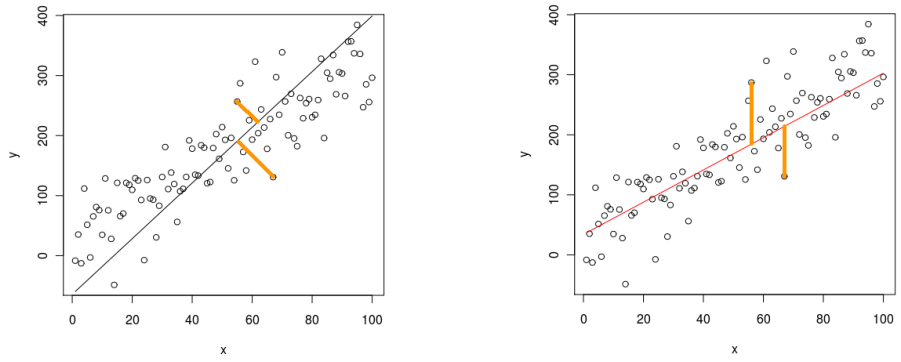

Figure 5: PCA (left-hand side) vs. OLS (right-hand side)[xxv)]

unequal[xxvi)]. Formally, let $x_1, \ldots, x_n \in \mathbb{R}^p$ be centered observations, and Π_q be the subspace spanned by the first q vectors of an orthonormal basis $f_1, \ldots, f_q, f_{q+1}, \ldots, f_p$. It follows that

$$\operatorname{proj}_{\Pi_q} x_i = \sum_{j=1}^{q} \langle x_i, f_j \rangle f_j.$$

If

$$d^2(x_i, \Pi_q) = \left\| x_i - \operatorname{proj}_{\Pi_q} x_i \right\|^2 = \sum_{j=q+1}^{p} \langle x_i, f_j \rangle^2,$$

then

$$\operatorname{span}(e_1, \ldots, e_q) = \underset{\Pi_q \subset \mathbb{R}^p : \dim(\Pi_q) = q}{\arg\min} \sum_{i=1}^{n} d^2(x_i, \Pi_q).$$

Proof. Denote by $S = \frac{1}{n-1} \sum_{i=1}^{n} x_i x_i'$ the sample covariance matrix. Direct calculations suggest that

$$\sum_{i=1}^{n} d^2(x_i, \Pi_q) = (n-1) \operatorname{tr}\left(S\big|_{\Pi_q^\perp} \right) = (n-1) \sum_{j=q+1}^{p} f_j' S f_j$$

attains its minimum if and only if $f_{q+1}, \ldots, f_p$ are the last $p - q$ eigenvectors of S. □

§18 Canonical Correlation & Factor Analysis (2019/12/10)

Given a random p-vector X and a random q-vector Y with $\operatorname{Var}\left(\begin{pmatrix} X \\ Y \end{pmatrix} \right) = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$, the **first pair of canonical variables** consists of $U_1 = e_1' X$ and $V_1 = f_1' Y$ such that $\operatorname{Corr}(U_1, V_1)$ is maximized among $e_1 \in \mathbb{R}^p$ and $f_1 \in \mathbb{R}^q$. Actually, $e_1 = \Sigma_{XX}^{-1/2} \tilde{e}_1$ and $f_1 = \Sigma_{YY}^{-1/2} \tilde{f}_1$, where $\tilde{e}_1$ and $\tilde{f}_1$ are the first left and right singular vectors of the **canonical correlation matrix** $C = \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}$. To see this, note that if $c = \Sigma_{XX}^{1/2} a$ for $a \in \mathbb{R}^p$ and $d = \Sigma_{YY}^{1/2} b$ for $b \in \mathbb{R}^q$, then

$$\operatorname{Corr}(a'X, b'Y) = \frac{a' \Sigma_{XY} b}{\sqrt{a' \Sigma_{XX} a} \sqrt{b' \Sigma_{YY} b}} = \frac{c' C d}{\|c\| \cdot \|d\|}.$$

Recall that the left and right singular vectors of C are eigenvectors of CC' and $C'C$, respectively. Sequentially, the k^{th} **pair of canonical variables** is $U_k = e_k' X$ and $V_k = f_k' Y$ that maximizes $\operatorname{Corr}(U_k, V_k)$ among those linear combinations uncorrelated with the preceding $k-1$ canonical pair(s). It turns out that $U_k = \tilde{e}_k' \Sigma_{XX}^{-1/2} X$ and $V_k = \tilde{f}_k' \Sigma_{YY}^{-1/2} Y$, where $\tilde{e}_k$ and $\tilde{f}_k$ are the k^{th} left and right singular vectors of C.

[xxv)] from JD Long's blog on 16 Sep 2010. https://cerebralmastication.com/
[xxvi)] cf. https://stats.stackexchange.com/q/22718

Factor analysis is widely used in social sciences for qualitative modeling, where observable/manifest variables $X = (X_1, \ldots, X_p)'$ are explained in terms of a potentially *smaller* number of underlying latent variables $F = (F_1, \ldots, F_m)'$ that are often impossible to measure directly. Such unobservable variables of interest are called **common factors**, because they influence the observed variables broadly. In psychology, *intelligence* is a prime example, which researchers examine indirectly by measuring variables that are believed to be its indicators, e.g., test scores in literature, language and mathematics. The essential purpose of factor analysis is to exhibit the relationship between the latent and the observed variables.

Let $\mu = \mathbb{E}X$. The classic **orthogonal factor model** is

$$\underset{(p\times 1)}{X-\mu} = \underset{(p\times m)(m\times 1)}{L}\underset{}{F} + \underset{(p\times 1)}{\varepsilon}, \quad \text{or} \quad X_i - \mu_i = \sum_{j=1}^{m} \ell_{ij}F_j + \varepsilon_i, \quad 1 \le i \le p$$

where ℓ_{ij} is called the **loading** of the i^{th} variable on the j^{th} factor, meaning the weight that the j^{th} factor carries in determining the i^{th} variable; and ε_i is the error (sometimes called the **specific factor**) associated only with the i^{th} variable. Note that the loading matrix $L = (\ell_{ij})$ does not vary across observations. Also, the common and specific factors are assumed to satisfy that

- $\mathbb{E}F = 0_m$, $\mathrm{Var}(F) = I_m$ to ensure that the factors are uncorrelated;
- $\mathbb{E}\varepsilon = 0_p$, $\mathrm{Var}(\varepsilon) = \Psi = \mathrm{diag}(\psi_1, \ldots, \psi_p)$ so that $\mathrm{Cov}(X_i, X_{i'}|F) = 0$ for $i \neq i'$; and
- $\mathrm{Cov}(F, \varepsilon) = 0_{m\times p}$.

Note that the assumptions still hold if we set $F \leftarrow Q'F$ and $L \leftarrow LQ$ for an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$, and thereby the model is identifiable only up to a *factor rotation*. From the assumptions, it follows the covariance structure $\underline{\mathrm{Var}(X) = LL' + \Psi}$, or

$$\mathrm{Var}(X_i) = \sum_{j=1}^{m} \ell_{ij}^2 + \psi_i, \quad \& \quad \mathrm{Cov}(X_i, X_{i'}) = \sum_{j=1}^{m} \ell_{ij}\ell_{i'j}, \quad \forall i \neq i'.$$

Here $h_i^2 = \sum_{j=1}^{m} \ell_{ij}^2$ is called the i^{th} **communality** and ψ_i is often called the **specific/unique variance**. It can also be seen that

$$\mathrm{Cov}(X, F) = \mathrm{Cov}(LF, F) + \mathrm{Cov}(\varepsilon, F) = L.$$

There is a strong connection between PCA and factor analysis (FA), but the two are not identical. Both methods are mostly used in exploratory data analysis for dimension reduction. However, PCA aims at explaining no covariances but the total variance, whereas FA is more concerned about accounting for the covariances or correlations. For comparison, write the PCA as

$$X - \mu = Q_1 Y^{(1)} + Q_2 Y^{(2)} = Q_1 Y^{(1)} + u,$$

where $(Q_1; Q_2) = (e_1, \ldots, e_m; e_{m+1}, \ldots, e_p) = (q_{ij})$ consists of eigenvectors of $\mathrm{Var}(X)$, and $Y^{(1)}$ and $Y^{(2)}$ are the first m and the last $p-m$ principal component(s) of X, respectively. Note that $\mathrm{Var}(u) = \sum_{j=m+1}^{p} \lambda_j e_j e_j'$ is usually not diagonal. It's difficult to interpret the i^{th} variable $X_i = \sum_{j=1}^{m} q_{ij}Y_j + u_i$. And we prefer the entire extraction $X = \sum_{j=1}^{m} Y_j e_j + u$. Indeed, PCA transforms observed variables to obtain principle components, while FA transforms factors to obtain observed variables. It's sensible to say how X_i's collectively form/interpret $Y_j = e_j'X$, rather than how Y_j's explain X_i.

The objective of factor analysis is to find the factor loadings L and the specific variances Ψ. In addition, we need to select the number of factors m. The first step is to estimate $\mathrm{Var}(X)$ by calculating the sample covariance matrix S, or the sample correlation matrix R, since the factor model is *scale invariant* in that scaled $(X_i/\sigma_i)_{1\le i\le p}$ admits the same latent factors. The estimates $\hat{L}$ of L and $\hat{\Psi}$ of Ψ are deduced from the covariance structure

$$\underset{\text{d.o.f.: } p(p+1)/2}{S} = \underset{pm}{\hat{L}\hat{L}'} + \underset{p}{\hat{\Psi}}.$$

Whether the model has a solution or not is determined by the degree of freedom $s = \frac{1}{2}p(p+1) - (pm+p)$.

- If $s < 0$, then the model is meaningless: the number of factors m is "too large" relative to the number of original parameters p and there are infinitely many solutions.

- If $s = 0$, then the solution is unique, but not necessarily proper — the specific variances Ψ are required to be non-negative.
- In practice we usually have $s > 0$: an exact solution may not exist and approximate solutions are used. This case is the most interesting for data/dimension reduction with interpretation.

There are two popular methods of estimation, the *principal component* (and the related *principal factor*) method and the *maximum likelihood* method.

Principal component method leads to an iterative solution as follows:

1. Calculate eigenpairs (λ_i, e_i) of the sample covariance matrix S;
2. Initialize $\hat{L} = (\sqrt{\lambda_1}e_1, \ldots, \sqrt{\lambda_m}e_m)$ and $\hat{\Psi} = \operatorname{diag}\left(S - \hat{L}\hat{L}'\right)$;
3. Update $\hat{L}$ using the eigen-decomposition of $S - \hat{\Psi}$;
4. Update $\hat{\Psi} = \operatorname{diag}\left(S - \hat{L}\hat{L}'\right)$;
5. Repeat steps 3. and 4. until convergence.

§19 Factor Analysis Cont'd & Classification (2019/12/12)

As an alternative for estimation of the factor model, **maximum likelihood method** assumes normality. Recall that

$$\log f(\mathbf{X}; \mu, \Sigma) = -\frac{n}{2}\left[\log(\det(2\pi\Sigma)) + \operatorname{tr}\left(\Sigma^{-1}S_n\right) + (\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu)\right],$$

where $S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})' = \frac{n-1}{n}S$. Replacing μ with $\hat{\mu} = \bar{X}$ and substituting $\Sigma = LL' + \Psi$, the estimates $\hat{L}$ and $\hat{\Psi}$ are obtained by numerical minimization of

$$\log(\det(LL' + \Psi)) + \operatorname{tr}\left(S_n(LL' + \Psi)^{-1}\right).$$

It is desirable to make $\hat{L}$ well defined by imposing the computationally convenient uniqueness condition (check!) that $L'\Psi^{-1}L$ is diagonal. MLE method has an *advantage*: one can test whether the number of factors m is adequate, i.e., Σ has the form $LL' + \Psi$ or not. The likelihood-ratio statistic is $-2\log(\Lambda) = nF$, where

$$F = \log\left(\frac{\det(\hat{L}\hat{L}' + \hat{\Psi})}{\det(S_n)}\right) + \operatorname{tr}\left(S_n(\hat{L}\hat{L}' + \hat{\Psi})^{-1}\right) - p.$$

Bartlett has suggested that

$$\left(n - 1 - \tfrac{1}{6}(2p + 5) - \tfrac{2}{3}m\right)F \xrightarrow[H_0]{d} \chi^2_\nu,$$

where the degree of freedom

$$\nu = \tfrac{1}{2}p(p+1) - \left[pm + p - \tfrac{1}{2}m(m-1)\right] = \tfrac{1}{2}\left[(p-m)^2 - (p+m)\right]$$

must be positive. In practice, we often start with the largest possible $m < \frac{1}{2}(2p + 1 - \sqrt{8p+1})$. The backward tests proceed as follows: if the adequacy of the m-factor model is not rejected, continue with $m - 1$ factors, until rejection. Also, forward testing procedure works.

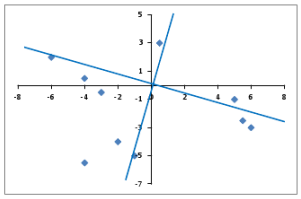

Figure 6: Orthogonal rotation

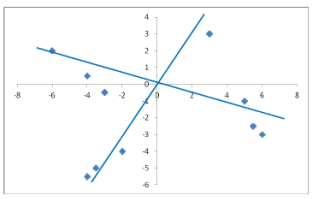

Figure 7: Oblique rotation

As is shown, an orthogonal transformation of factors, referred to as **factor rotation**, preserves the ability to reproduce the covariance. We can utilize the non-uniqueness property to deliberately change the loadings in order to improve their interpretability. Figures 6 & 7 are visuals[xxvii] of what happens during

[xxvii] from https://www.theanalysisfactor.com/rotations-factor-analysis/

a rotation of two factors. Note that oblique rotations allow for correlated factors. The loadings can be interpreted easily if a "simple clusters structure" is achieved — each variable is highly loaded on a single factor and little on others. Given the estimated factor loadings $\hat{L} = (\hat{\ell}_{ij})$, Kaiser proposed an analytical measure of simple structure known as the **varimax criterion**. The idea of the *varimax rotation method* is to find the rotation that maximizes the variances of the squared loadings within each column since each factor should have a few large and many negligible loadings. More precisely, the varimax criterion of $L = (\ell_{ij})$ is defined to be

$$\mathrm{VC}(L) = \sum_{j=1}^{m} \left[\frac{1}{p} \sum_{i=1}^{p} \widetilde{\ell}_{ij}^4 - \left(\frac{1}{p} \sum_{i=1}^{p} \widetilde{\ell}_{ij}^2 \right)^2 \right],$$

where $\widetilde{\ell}_{ij} = \ell_{ij}/\hat{h}_i$ using the square root of the estimated i^{th} communality $\hat{h}_i^2 = \sum_{j=1}^{m} \hat{\ell}_{ij}^2$. The varimax optimal rotation is chosen so that $\mathrm{VC}(\hat{L}Q)$ is maximized among orthogonal matrices $Q \in \mathbb{R}^{m \times m}$.

The aim of **discriminant analysis** is to classify objects into groups that are known *a priori*. Assume that data come from g distinct classes/populations $D_1, \ldots, D_g$. Given an observed random vector x, which population should *have* generated x? Denote by $f_j(x)$ the densities of each population D_j, respectively. The **maximum likelihood principle** is implemented by allocating x to D_J so that

$$J = \underset{j \in \{1, \cdots, g\}}{\arg\max}\ f_j(x).$$

Consider $g = 2$ univariate normal populations $D_1 = \mathcal{N}(\mu_1, \sigma_1^2)$ and $D_2 = \mathcal{N}(\mu_2, \sigma_2^2)$. Then

$$\begin{aligned} f_1(x) > f_2(x) &\iff \frac{1}{\sigma_1} \exp\left\{ -\frac{(x-\mu_1)^2}{2\sigma_1^2} \right\} > \frac{1}{\sigma_2} \exp\left\{ -\frac{(x-\mu_2)^2}{2\sigma_2^2} \right\} \\ &\iff \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right) x^2 - 2 \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \right) x + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} \right) < 2 \log \left(\frac{\sigma_2}{\sigma_1} \right), \end{aligned}$$

which simplifies to

$$(\mu_1 - \mu_2) \left(x - \frac{\mu_1 + \mu_2}{2} \right) > 0$$

provided that $\sigma_1 = \sigma_2$.

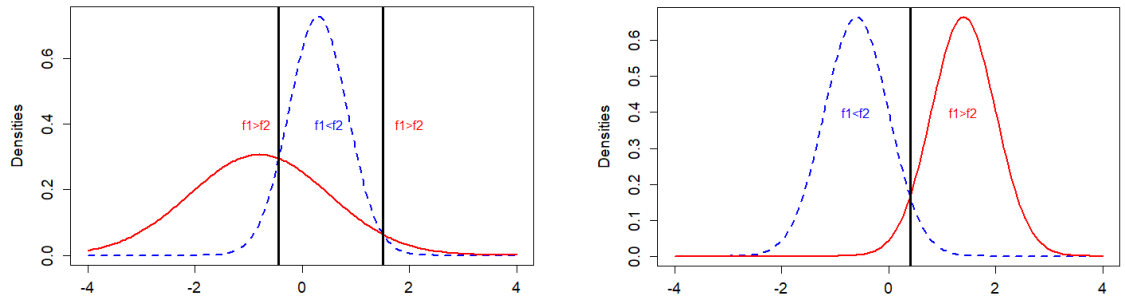

Figure 8: Maximum likelihood rule for univariate normal distributions[xxviii)]

In multivariate setting, suppose $D_j = \mathcal{N}_p(\mu_j, \Sigma)$. Note that the covariance matrices are *equal* for all classes. Then

$$\begin{aligned} f_1(x) > f_2(x) &\iff (x-\mu_1)'\Sigma^{-1}(x-\mu_1) < (x-\mu_2)'\Sigma^{-1}(x-\mu_2) \\ &\iff \log\left(\tfrac{f_1(x)}{f_2(x)} \right) = (\mu_1 - \mu_2)'\Sigma^{-1} \left(x - \tfrac{1}{2}(\mu_1 + \mu_2) \right) > 0. \end{aligned}$$

Note that $\|x - \mu\|_\Sigma = \sqrt{(x-\mu)'\Sigma^{-1}(x-\mu)}$ is called the **Mahalanobis distance**. The classification rule based on linear combinations belongs to the family of **linear discriminant analysis** (LDA) methods.

§20 Classification Cont'd (2019/12/17)

The *decision boundary* $\left\{ x \in \mathbb{R}^p : (\mu_1 - \mu_2)'\Sigma^{-1} \left(x - \frac{1}{2}(\mu_1 + \mu_2) \right) = 0 \right\}$ is often indicated in contour plots, as a hyperplane that goes through the point $\frac{1}{2}(\mu_1 + \mu_2)$ and has normal vector $a = \Sigma^{-1}(\mu_1 - \mu_2)$.

[xxviii)] generated by https://github.com/QuantLet/MVA/blob/master/QID-1209-MVAdisnorm/MVAdisnorm.r

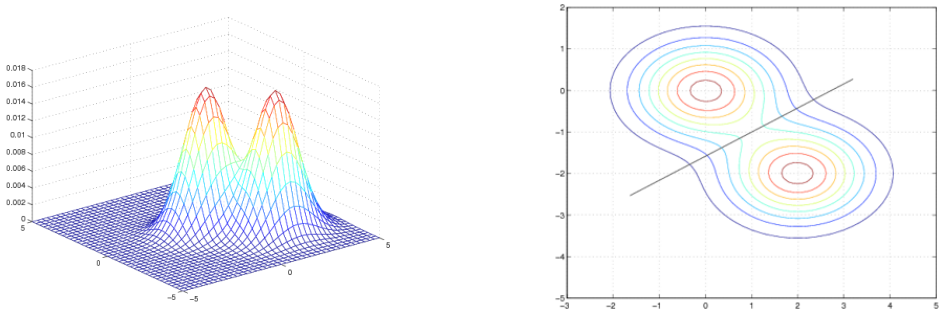

Figure 9: LDA decision boundary for bivariate normal distributions[xxix)]

For $D_j = \mathcal{N}_p(\mu_j, \Sigma_j)$ with $\Sigma_1 \neq \Sigma_2$, we have $f_1(x) > f_2(x)$ if and only if

$$\log\left(\frac{f_1(x)}{f_2(x)}\right) = -\frac{1}{2}x'(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})x - \left[\frac{1}{2}\log\left(\frac{\det\Sigma_1}{\det\Sigma_2}\right) + \frac{1}{2}\left(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2\right)\right] > 0.$$

This classification rule belongs to the family of **quadratic discriminant analysis** (QDA) methods.

Denote the *classifier* by G, i.e.,

$$G = j \iff X \sim D_j.$$

In Bayesian view, past experience gives rise to *prior probabilities*

$$\pi_j = p(D_j) = \mathbb{P}(G = j), \quad j \in \{1, \cdots, g\}$$

before data is collected. Then, Bayes' theorem allows us to update our previous beliefs with new information. Maximization of the *posterior probability*

$$p(D_j|x) = \mathbb{P}(G = j|x) = \frac{\pi_j f_j(x)}{\sum_{k=1}^g \pi_k f_k(x)} \propto \pi_j f_j(x)$$

leads to the **naïve Bayesian classifier** (see also Lecture §3)

$$\hat{G}(x) = \underset{j\in\{1,\cdots,g\}}{\arg\max}\ p(D_j|x) = \underset{j\in\{1,\cdots,g\}}{\arg\max}\ \pi_j f_j(x).$$

Note that

$$\pi_1 f_1(x) > \pi_2 f_2(x) \iff \log\left(\frac{f_1(x)}{f_2(x)}\right) > \log\left(\frac{\pi_2}{\pi_1}\right).$$

The Bayes rule of discrimination will coincide with the maximum likelihood principle if, $\pi_j \propto 1$ correspond to the uniform distribution, i.e., the prior is *non-informative*.

To apply these methods, one gathers a *training sample*, for which the correct classes are known (*labeled*). Then it suffices to estimate parameters as follows:

- $\mu_j \leftarrow \bar{x}_j$, sample mean of observations that belong to class j.
- $\Sigma \leftarrow S_{\text{pooled}}$, sample covariance matrix of pooled observations, applicable to LDA.
- $\Sigma_j \leftarrow S_j$, sample covariance matrix of observations that belong to class j, applicable to QDA.
- $\pi_j \leftarrow n_j/n$, proportion of observations in the data that belong to class j.

We may encounter a misclassification error when the observation is assigned to a certain group. Denote by $A_j \subset \mathbb{R}^p$ the **classification region** (a.k.a. the *action space* in *statistical decision theory*) such that we allocate $x \in A_j$ to D_j. Let $c(k|j) \in \mathbb{R}_+$ be the *cost* when an observation actually from D_j falls into A_k and is mistakenly assigned to D_k. The probability of putting $X \sim D_j$ into D_k can be calculated as

$$p(k|j) = \mathbb{P}(X \in A_k|D_j) = \int_{A_k} f_j(x)\,\mathrm{d}x.$$

[xxix)] from https://www.personal.psu.edu/jol2/course/stat597e/notes2/lda.pdf

Suppose there are only two populations. The **expected cost of misclassification** is then

$$\mathrm{ECM} = c(2|1)p(2|1)\pi_1 + c(1|2)p(1|2)\pi_2.$$

The rule minimizing the ECM among all partitions $A_1 \sqcup A_2 = \mathbb{R}^p$ is given by (check!)

$$x \in A_1 \iff c(2|1)\pi_1 f_1(x) > c(1|2)\pi_2 f_2(x) \iff \log\left(\frac{f_1(x)}{f_2(x)}\right) > \log\left(\frac{c(1|2)\pi_2}{c(2|1)\pi_1}\right).$$

This is exactly the Bayes discrimination rule as long as the misclassification costs are equal. Thus, the Bayes rule is optimal in that it minimizes the **total probability of misclassification**

$$\mathrm{TPM} = p(2|1)\pi_1 + p(1|2)\pi_2.$$

Note that if two or more classification rules are available, it is important to evaluate the performance of each rule. The smallest possible TPM is called the **optimum error rate** (OER). Let us derive an expression for the OER when $\pi_1 = \pi_2 = \frac{1}{2}$ and $D_j = \mathcal{N}_p(\mu_j, \Sigma)$. LDA yields

$$A_1 = \left\{x \in \mathbb{R}^p : a'x > \tfrac{1}{2}a'(\mu_1+\mu_2)\right\}, \quad \& \quad A_2 = \left\{x \in \mathbb{R}^p : a'x \le \tfrac{1}{2}a'(\mu_1+\mu_2)\right\},$$

where $a = \Sigma^{-1}(\mu_1 - \mu_2)$. After tedious calculations, one can see that

$$p(2|1) = p(1|2) = \Phi(-\Delta/2) \implies \mathrm{OER} = \Phi(-\Delta/2),$$

where $\Delta = \sqrt{a'\Sigma a} = \|\mu_1 - \mu_2\|_\Sigma$. Generally speaking, estimated TPM is bitterly subtle and involved.

§21 Classification Cont'd (2019/12/24)

The performance of classifiers can be empirically evaluated by calculating the estimate of TPM, i.e.,

$$\widehat{\mathrm{TPM}} = \hat\pi_1\hat p(2|1) + \hat\pi_2\hat p(1|2) = \frac{n_1}{n}\int_{\hat A_2}\hat f_1(x)\,\mathrm{d}x + \frac{n_2}{n}\int_{\hat A_1}\hat f_2(x)\,\mathrm{d}x$$

using a sample of size $n = n_1 + n_2$.

The probabilities of misclassification may also be estimated by the *re-substitution method*, which does not depend on the form of the parent populations. We reclassify the *testing* sample according to a rule obtained from the *training* sample. Then we have $\hat p(k|j) = n_{jk}/n_j$ as an estimate of $p(k|j)$, where n_{jk} is the number of individuals coming actually from D_j that are classified into D_k.

		Estimated class 1	Estimated class 2	
True	class 1	n_{11}	n_{12}	n_1
	class 2	n_{21}	n_{22}	n_2

Table 4: Confusion matrix in binary classification

The estimated TPM is called **error rate** here, defined as

$$\widehat{\mathrm{TPM}} = \sum_{j=1}^{g}\sum_{k\neq j}\hat p(k|j)\hat\pi_j = \sum_{j=1}^{g}\sum_{k\neq j}\frac{n_{jk}}{n_j}\frac{n_j}{n} = 1 - \frac{1}{n}\sum_{j=1}^{g} n_{jj}.$$

If all samples are used to train a classifier, then $\widehat{\mathrm{TPM}}_{\mathrm{all}}$ tends to be accompanied by downward bias, which means that the classification rule appears overly optimistic.

An approach that corrects for this bias is based on the *holdout procedure*. The original observations are split into a *training* sample and a *testing* sample recurrently, and the cross-validation (CV, see also Lecture §16) with the 0-1 loss applies. The leave-one-out method, which is a special case of K-fold CV, is frequently used and often results in an unbiased $\widehat{\mathrm{TPM}}_{\mathrm{CV}}$. Thereby, we usually have $\widehat{\mathrm{TPM}}_{\mathrm{CV}} > \widehat{\mathrm{TPM}}_{\mathrm{all}}$. For example, see exercise 2 in homework 5.

By the way, data are often partitioned into three parts if there exist tuning parameters. The *tuning sample* is used to numerically assess and determine parameters first. The *training* and *testing* sample are then used to compare the proposed method with others.

Fisher's discriminant analysis is a distance-based method, which is extensively applied due to its geometrical interpretation. The idea is to find a linear rule (in other words, base the classification rule on a *projection*, say, $x \in \mathbb{R}^p \mapsto a'x \in \mathbb{R}$) that separates the classes best. Denote the observations by $x_{ij} \in \mathbb{R}^p$, $i = 1, \cdots, n_j$; $j = 1, \cdots, g$. Let $y_{ij} = a'x_{ij}$ for some $a \in \mathbb{R}^p$. In the case when $g = 2$, the objective is

$$\max_{a \in \mathbb{R}^p} \frac{|\bar{y}_{\cdot 1} - \bar{y}_{\cdot 2}|}{\sqrt{s_y^2}}, \quad \text{where } s_y^2 = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (y_{i1} - \bar{y}_{\cdot 1})^2 + \sum_{i=1}^{n_2} (y_{i2} - \bar{y}_{\cdot 2})^2 \right).$$

Note that Fisher's approach does not assume that the populations are normal. It does, however, implicitly assume that the population covariance matrices are equal. Write

$$S_x = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_{\cdot 1})(x_{i1} - \bar{x}_{\cdot 1})' + \sum_{i=1}^{n_2} (x_{i2} - \bar{x}_{\cdot 2})(x_{i2} - \bar{x}_{\cdot 2})' \right).$$

Then

$$\frac{|\bar{y}_{\cdot 1} - \bar{y}_{\cdot 2}|}{\sqrt{s_y^2}} = \frac{|a'(\bar{x}_{\cdot 1} - \bar{x}_{\cdot 2})|}{\sqrt{a' S_x a}} \le \sqrt{(\bar{x}_{\cdot 1} - \bar{x}_{\cdot 2})' S_x^{-1} (\bar{x}_{\cdot 1} - \bar{x}_{\cdot 2})} = \|\bar{x}_{\cdot 1} - \bar{x}_{\cdot 2}\|_{S_x}$$

by Cauchy-Schwarz inequality, with equality holding if and only if $a \propto \hat{a} = S_x^{-1}(\bar{x}_{\cdot 1} - \bar{x}_{\cdot 2})$, which has

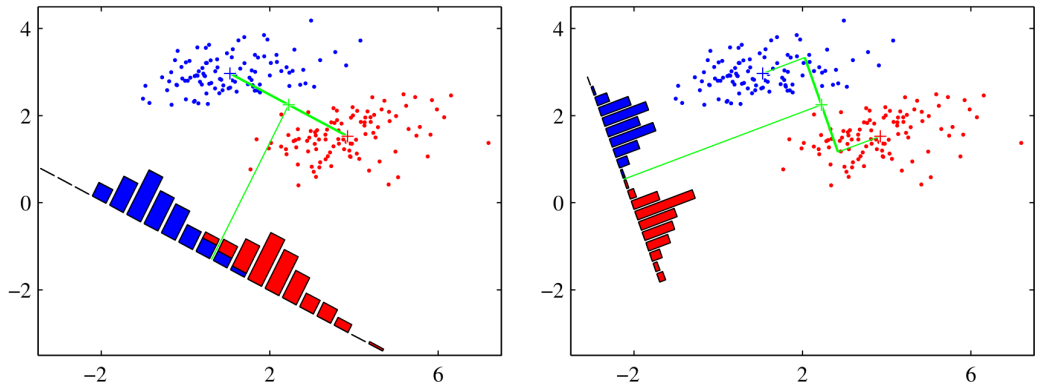

Figure 4.6 The left plot shows samples from two classes (depicted in red and blue) along with the histograms resulting from projection onto the line joining the class means. Note that there is considerable class overlap in the projected space. The right plot shows the corresponding projection based on the Fisher linear discriminant, showing the greatly improved class separation.

Figure 10: Fisher's linear discriminant[xxx)]

already occurred in Gaussian LDA. An observation $x \in \mathbb{R}^p$ will be allocated to D_1 if and only if

$$|\hat{a}'(x - \bar{x}_{\cdot 1})| < |\hat{a}'(x - \bar{x}_{\cdot 2})| \iff \hat{a}' \left(x - \tfrac{1}{2}(\bar{x}_{\cdot 1} + \bar{x}_{\cdot 2}) \right) > 0.$$

Generally, we seek $a \in \mathbb{R}^p$ to maximize the ratio of the *between-group variance* to the *within-group variance*, for g classes with equal covariance matrices. Let (see LECTURE §12, table 1)

$$B = \sum_{j=1}^{g} n_j (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})', \quad \& \quad W = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{\cdot j})(x_{ij} - \bar{x}_{\cdot j})'.$$

Note that $S_{\text{pooled}} = W/(n - g)$. Then,

$$\sum_{j=1}^{g} n_j (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 = a'Ba, \quad \& \quad \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\cdot j})^2 = a'Wa.$$

[xxx)] from Bishop's PRML; see also https://stackoverflow.com/q/33844198

One can show that

$$\arg\max_{a\in\mathbb{R}^p} \frac{a'Ba}{a'Wa} \overset{(b=W^{1/2}a)}{=\!=\!=} \frac{b'W^{-1/2}BW^{-1/2}b}{b'b}$$

is given (check!) by the first eigenvector $\hat{a}_1$ of $W^{-1}B$, producing the **first Fisher's sample discriminant**. Now a classification rule is easy to obtain: we allocate $x \in \mathbb{R}^p$ to class J_1 where

$$J_1 = \arg\min_{j\in\{1,\cdots,g\}} \left|\hat{a}_1'(x-\bar{x}_{\cdot j})\right|.$$

We can utilize more linear discriminants in order to summarize the distinction between the classes in higher dimension. Let

$$\hat{a}_k = \arg\max_{a\in\mathbb{R}^p} a'Ba \quad \text{s.t.} \quad a'S_{\text{pooled}}a = 1;\ a'S_{\text{pooled}}\hat{a}_\ell = 0,\ \forall \ell < k.$$

It turns out that $\hat{a}_k$ is the k^{th} eigenvector of $S_{\text{pooled}}^{-1}B$ for $1 \le k \le p$. The number of $\hat{a}_k$'s corresponding to nonzero eigenvalues is

$$s = \operatorname{rank}\left(W^{-1}B\right) \le \min(p, g-1),$$

since B is spanned by those $\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}$ such that $\sum_{j=1}^g n_j(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) = 0_p$. Also, for any $k > s$,

$$B\hat{a}_k = 0_p \implies \hat{a}_k'B\hat{a}_k = 0 \implies \hat{a}_k'(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) = 0,\ \forall j = 1,\cdots,g.$$

Allocate $x \in \mathbb{R}^p$ to class

$$J_r = \arg\min_{j\in\{1,\cdots,g\}} \sum_{k=1}^r \left|\hat{a}_k'(x-\bar{x}_{\cdot j})\right|^2.$$

Note that

$$\sum_{k=1}^p \left|\hat{a}_k'(x-\bar{x}_{\cdot j})\right|^2 = (x-\bar{x}_{\cdot j})'\underbrace{\left(\sum_{k=1}^p \hat{a}_k\hat{a}_k'\right)}_{=S_{\text{pooled}}^{-1}}(x-\bar{x}_{\cdot j}) = \|x-\bar{x}_{\cdot j}\|^2_{S_{\text{pooled}}}.$$

If $r < s$ discriminants are used for classification, there is a loss of squared distance of $\sum_{k=r+1}^p \left|\hat{a}_k'(x-\bar{x}_{\cdot j})\right|^2$, where $\sum_{k=r+1}^s \left|\hat{a}_k'(x-\bar{x}_{\cdot j})\right|^2$ is the part useful for classification. Note that $\hat{a}_k'(\bar{x}_{\cdot j'} - \bar{x}_{\cdot j}) = 0$ for $k > s$, and therefore $\sum_{k=s+1}^p \left|\hat{a}_k'(x-\bar{x}_{\cdot j})\right|^2$ is constant — does not depend on j.

§22 Classification Cont'd & Final Review (2019/12/26)

If prior probabilities π_j are considered, we may allocate $x \in \mathbb{R}^p$ to class

$$\arg\min_{j\in\{1,\cdots,g\}} \left\{\sum_{k=1}^r \left|\hat{a}_k'(x-\bar{x}_{\cdot j})\right|^2 - 2\log(\pi_j)\right\},$$

or equivalently,

$$\arg\max_{j\in\{1,\cdots,g\}} \left\{-\frac{1}{2}\sum_{k=1}^r \left|\hat{a}_k'(x-\bar{x}_{\cdot j})\right|^2 + \log(\pi_j)\right\}.$$

Assume that there are only two classes, indexed by an indicator Y that takes the value from $\{0,1\}$. In **logistic regression**, one directly models

$$p(x) := \mathbb{P}(Y=1|X=x) = \mathbb{E}[Y|X=x].$$

The logarithm of the *odds* $\frac{p}{1-p}$ is called the **logit** of the probability p, i.e., $\text{logit} : p \in [0,1] \mapsto \log\left(\frac{p}{1-p}\right) \in \mathbb{R}$. It is a special case of a **link function** in a *generalized linear model* (GLM[xxxi)]) — for the Bernoulli distribution. As in linear regression,

$$\text{logit}(p(x)) = \alpha + x'\beta.$$

[xxxi)] see https://zhuanlan.zhihu.com/p/80178516 for a brief (p)review. More generally, when an exact distribution is difficult to handle, we often appeal to corresponding large sample properties.

Then

$$p(x) = \frac{\exp(\alpha + x'\beta)}{1 + \exp(\alpha + x'\beta)}.$$

It follows that the log-likelihood function is

$$\ell_n(\alpha, \beta) = \log\left(\prod_{i=1}^n [p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i}\right) = \sum_{i=1}^n \Big[(\alpha + x_i'\beta)y_i - \log\left(1 + \exp(\alpha + x_i'\beta)\right)\Big].$$

To obtain the MLE for $\theta = (\alpha, \beta)$, we often solve

$$\begin{cases} \dfrac{\partial \ell_n}{\partial \alpha} = \sum\limits_{i=1}^n \left(y_i - \frac{\exp(\alpha + x_i'\beta)}{1+\exp(\alpha + x_i'\beta)}\right) & = 0 \\ \dfrac{\partial \ell_n}{\partial \beta} = \sum\limits_{i=1}^n \left(y_i - \frac{\exp(\alpha + x_i'\beta)}{1+\exp(\alpha + x_i'\beta)}\right) x_i & = 0_p \end{cases}$$

numerically; for instance, the *Newton–Raphson method* is iterated as

$$\hat\theta^{(t+1)} = \hat\theta^{(t)} - \ddot\ell_n(\hat\theta^{(t)})^{-1}\dot\ell_n(\hat\theta^{(t)}).$$

Note that $\ddot\ell_n(\theta)$ depends on θ and x_i's, but not on y_i's. In contrast to LDA, logistic regression relies on *conditional likelihood* and requires a large sample size, which means that logistic regression is less efficient. LDA, relying only on the equal covariance matrices assumption, is less robust.

Nonparametric classification is based mainly on **density estimation**. A generalization of the kNN estimate (see LECTURE §3) is given by $\hat p_n(x) = \frac{1}{nh}\sum_{i=1}^n K(\frac{x_i - x}{h})$, called the **kernel density estimate**, where K is a *kernel* and h is the *bandwidth*.

As for the FINAL EXAM, important topics are

- ◇ eigen-decomposition and SVD (see LECTURE 1,2)
- ◇ multivariate normal distr — conditional distr, normal matrix, Wishart distr (see LECTURE 5,6,8,9)
- ◇ inference about means — one sample, two samples (see LECTURE 9–11)
- ◇ Cochran's Theorem — sum of squares decomposition (see LECTURE 9)
- ◇ multivariate linear regression — basics (see LECTURE 13,14)
- ◇ PCA and FA — statistical interpretation (see LECTURE 16–18)
- ◇ classification — LDA applied to two populations (see LECTURE 19–21)

The End