Important Misc in Probab/Stat[1]

1 Important matrix decompositions

1. **Eigen Decomposition**: Let $\mathbf{P}$ be a matrix of eigenvectors of a given square matrix $\mathbf{A}$ and $\mathbf{D}$ be a diagonal matrix with the corresponding eigenvalues on the diagonal. Then, as long as $\mathbf{P}$ is a square matrix, $\mathbf{A}$ can be written as an eigen decomposition
$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1},$$
where $\mathbf{D}$ is a diagonal matrix. Furthermore, if $\mathbf{A}$ is symmetric, then the columns of $\mathbf{P}$ are orthogonal vectors.
If $\mathbf{P}$ is not a square matrix (for example, the space of eigenvectors of $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is one-dimensional), then $\mathbf{P}$ cannot have a matrix inverse and $\mathbf{A}$ does not have an eigen decomposition. However, if $\mathbf{P}$ is $m \times n$ (with $m > n$), then $\mathbf{A}$ can be written using a so-called singular value decomposition.
2. **QR-decomposition** (Gram-Schmidt orthogonality): For any matrix $\boldsymbol{A}_{n\times m}$, there exists QR-decomposition $\boldsymbol{A} = \boldsymbol{Q}_{n\times m}\boldsymbol{R}_{m\times m}$, where $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_{m\times m}$ and $\boldsymbol{R}$ is an upper triangular.
3. **QR-decomposition** (Another version) For any matrix $\boldsymbol{A}_{n\times m}$ of rank k, there exists QR-decomposition $\boldsymbol{A} = \boldsymbol{Q}_{n\times m}\boldsymbol{R}_{m\times m}$, where $\boldsymbol{Q}^\top\boldsymbol{Q}$ is diagonal and $\boldsymbol{R}$ is a **unit** upper triangular.
4. **Left orthogonal decomposition** For any matrix $\boldsymbol{A}_{n\times m}$, there exist non-singular matrix $\boldsymbol{P}_{m\times m}$ and orthogonal matrix $\boldsymbol{G}_{n\times n}$ such that $\boldsymbol{A} = \boldsymbol{G}\begin{bmatrix} \boldsymbol{I}_r & 0 \\ 0 & 0 \end{bmatrix}\boldsymbol{P}$, where $r = rank(\boldsymbol{A})$.
5. **Cholesky Decomposition**: For positive matrix $\boldsymbol{A}_{n\times n}$, there exists Cholesky-decomposition $\boldsymbol{A} = \boldsymbol{T}^\top\boldsymbol{T}$, where $\boldsymbol{T}$ is an upper triangular. Also $\boldsymbol{T}$ is unique.
6. $\boldsymbol{A}$ and $\boldsymbol{B}$ are real symmetric matrices. Then there exists a orthogonal matrix P such that $P^\top\boldsymbol{A}P$ and $P^\top\boldsymbol{B}P$ are both diagonal if and only if $\boldsymbol{AB} = \boldsymbol{BA}$.
7. **Spectral decomposition** (A special case of item 1): Let $\boldsymbol{A}$ be $n\times n$ symmetric matrix. There exists an orthogonal matrix $\boldsymbol{T} = (t_1, \cdots, t_n)$ such that $\boldsymbol{T}^\top\boldsymbol{A}\boldsymbol{T} = \text{diag}(\lambda_1, \cdots, \lambda_n) = \Lambda$, where $\lambda_1 \geq \cdots \geq \lambda_n$ are the ordered eigenvalues of $\boldsymbol{A}$. With this ordering, Λ is unique and $\boldsymbol{T}$ is unique up to a postfactor.
$$\boldsymbol{A} = \sum_{i=1}^n \lambda_i t_i t_i^\top.$$

2 Normal Distribution

- If
$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} \sim \text{Normal}\left\{\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}\right\}.$$
Then $\mathbf{Y}|\mathbf{X} \sim \text{Normal}\left(\mu_{y|x}, \Sigma_{y|x}\right)$ with $\mu_{y|x} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\mathbf{X} - \mu_x)$ and $\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$

- A useful equality (Φ, ϕ: normal CDF, pdf)
$$\int \Phi(a+bx)\phi(x)dx = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right).$$
Proof. Let X, Z be iid random variables following standard normal. Then $\Phi(a+bx) = Pr(Z \leq a+bx)$. Note that
$$\begin{aligned}\int \Phi(a+bx)\phi(x)dx &= E_X\Phi(a+bX) \\ &= E_X\{P_Z(Z \leq a+bX)\} \\ &= E_X\{P_Z(Z - bX \leq a)\} \\ &= P_{(Z,X)}(Z - bX \leq a) \\ &= P\left\{N(0,1) \leq a/\sqrt{1+b^2}\right\} = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)\end{aligned}$$

- **Stein formulae**: X is a $N(0,1)$ random variable and g is an indefinite integral of the Lebesgue measurable function such that $E|g'(X)| < \infty$. Then
$$E\{g'(X)\} = E\{Xg(X)\}.$$

3 Linear and Quadratic Forms

1. $q = \mathbf{y}^\top\boldsymbol{A}\mathbf{y}$ is called a quadratic form in $\mathbf{y}$. $E(q) = \text{tr}(\boldsymbol{AV}) + \mu^\top\boldsymbol{A}\mu$ ($\mathbf{y}$ may not be normal) and $\text{cov}(q_1, q_2) = 2\text{tr}(\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_2\boldsymbol{V}) + 4\mu^\top\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_2\mu$.
2. $\mathbf{y} \sim N(\mu, \boldsymbol{V})$ then its characteristic function is $m_\mathbf{y}(t) = \exp\{t^\top\mu + \frac{1}{2}t^\top\boldsymbol{V}t\}$. The conditional density of $\mathbf{y_2}$ given $\mathbf{y_1}$ is $N(\mu_2 + \boldsymbol{V}_{21}\boldsymbol{V}_{11}^{-1}(y_1 - \mu_1), \boldsymbol{V}_{22} - \boldsymbol{V}_{21}\boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{12})$.
3. **Craig's Theorem** $\mathbf{y} \sim N(\mu, \boldsymbol{V})$. $\boldsymbol{B}\mathbf{y}$ is independent of $\mathbf{y}^\top\boldsymbol{A}\mathbf{y}$ iff $\boldsymbol{BVA} = 0$; $\mathbf{y}^\top\boldsymbol{B}\mathbf{y}$ is independent of $\mathbf{y}^\top\boldsymbol{A}\mathbf{y}$ iff $\boldsymbol{BVA} = 0$.
$q = \mathbf{y}^\top\boldsymbol{A}\mathbf{y} \sim \mathcal{X}^2(r, \lambda)$, with $r = r(\boldsymbol{A})$ and $\lambda = 1/2\mu^\top\boldsymbol{A}\mu$, if and only if $\boldsymbol{AV}$ is idempotent.
4. **Cochran's Theorem** $r(\boldsymbol{A}_i) = r_i$. Let $\boldsymbol{A} = \sum_1^k \boldsymbol{A}_i$ if $\boldsymbol{AV}$ is idempotent and $r(\boldsymbol{A}) = \sum_1^k r_i$, then $q_i = \mathbf{y}^\top\boldsymbol{A}_i\mathbf{y}$ are mutually independent, noncentral chi-squared variables with $\mathcal{X}^2(r_i, 1/2\mu\boldsymbol{A}_i\mu)$.

4 Algebra

algebra.tex

$\boldsymbol{A} = (a_{ij})_{sn}$, $\boldsymbol{B} = (b_{ij})_{nm}$; $\text{tr}(\boldsymbol{A})$: the trace of $\boldsymbol{A}$; $r(\boldsymbol{A})$: the rank of $\boldsymbol{A}$; $\det(\boldsymbol{A})$: the determinant of $\boldsymbol{A}$.

4.1 Trace and Eigenvalues

1. $\text{tr}(\boldsymbol{A}+\boldsymbol{B}) = \text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B})$.
2. $\text{tr}(\boldsymbol{AB}) = \text{tr}(\boldsymbol{BA})$.
3. $\boldsymbol{A}$ and $\boldsymbol{B}$ are real symmetric. Then $\text{tr}(\boldsymbol{ABAB}) \leq \text{tr}(\boldsymbol{A}^2\boldsymbol{B}^2)$, the equality holds if and only if $\boldsymbol{AB} = \boldsymbol{BA}$.
4. If $\boldsymbol{A}_n$ is a symmetric and $r(\boldsymbol{A}) = 1$, then $|I_n + \boldsymbol{A}_n| = 1 + \text{tr}(\boldsymbol{A})$.
5. For any $n \times n$ matrix $\boldsymbol{A}$ with eigenvalues $\lambda_1, \cdots, \lambda_n$, we have the following:
   (a) $\text{tr}(\boldsymbol{A}) = \sum\lambda_i$
   (b) $\det(\boldsymbol{A}) = \prod\lambda_i$
   (c) $\det(I_n \pm \boldsymbol{A}) = \prod(1 \pm \lambda_i)$
6. For conformable matrices, the nonzero eigenvalues of $\boldsymbol{AB}$ are the same as those of $\boldsymbol{BA}$.

4.2 Rank

1. $r(\boldsymbol{AB}) \geq r(\boldsymbol{A}) + r(\boldsymbol{B}) - n$.
2. $\boldsymbol{A} = (a_{ij})_{nn}$. If $\boldsymbol{A}^2 = I_n$. Then $r(\boldsymbol{A} + I_n) + r(\boldsymbol{A} - I_n) = n$.
3. $r(\boldsymbol{A}+\boldsymbol{B}) \leq r(\boldsymbol{A}) + r(\boldsymbol{B})$.
4. If $\boldsymbol{AB} = 0$. $r(\boldsymbol{A}) + r(\boldsymbol{B}) \leq n$.
5. $r(\boldsymbol{A}) = r(\boldsymbol{A}^\top\boldsymbol{A}) = r(\boldsymbol{A}\boldsymbol{A}^\top)$.
6. If $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are $m\times n$, $n \times p$ $p \times q$ matrices, then $r(\boldsymbol{AB}) + r(\boldsymbol{BC}) \leq r(\boldsymbol{B}) + r(\boldsymbol{ABC})$.

4.3 Patterned Matrices

1. If $\boldsymbol{A}$ and $\boldsymbol{C}$ are symmetric and all inverses exist,
$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} + \boldsymbol{F}\boldsymbol{E}^{-1}\boldsymbol{F}^\top & -\boldsymbol{F}\boldsymbol{E}^{-1} \\ -\boldsymbol{E}^{-1}\boldsymbol{F}^\top & \boldsymbol{E}^{-1} \end{pmatrix}$$
where $\boldsymbol{E} = \boldsymbol{C} - \boldsymbol{B}^\top\boldsymbol{A}^{-1}\boldsymbol{B}$ and $\boldsymbol{F} = \boldsymbol{A}^{-1}\boldsymbol{B}$.
2. $$\begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{vmatrix} = \begin{cases} |\boldsymbol{D}||\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}| & \text{if } \boldsymbol{D}^{-1} \text{ exists,} \\ |\boldsymbol{A}||\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}| & \text{if } \boldsymbol{A}^{-1} \text{ exists.} \end{cases}$$
Proof.
$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{pmatrix} \cdot \begin{pmatrix} I_r & O \\ -\boldsymbol{D}^{-1}\boldsymbol{C} & I_s \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{B} \\ O & \boldsymbol{D} \end{pmatrix}.$$
3. For matrixes $\boldsymbol{B}_{n\times m}$ and $\boldsymbol{C}_{m\times n}$, and non-singular $\boldsymbol{A}_{n\times n}$, $|\boldsymbol{A} + \boldsymbol{BC}| = |\boldsymbol{A}||\boldsymbol{I}_m + \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}|$.
4. $(\boldsymbol{I} + \boldsymbol{AB})^{-1} = \boldsymbol{I} - \boldsymbol{A}(\boldsymbol{I} + \boldsymbol{BA})^{-1}\boldsymbol{B}$; $|\boldsymbol{I} + \boldsymbol{AB}| = |\boldsymbol{I} + \boldsymbol{BA}|$; $|\boldsymbol{A}| = |\boldsymbol{A}_{11}||\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}|$.
5. $(\boldsymbol{A} + \boldsymbol{UBV})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{UB}(\boldsymbol{B} + \boldsymbol{BVA}^{-1}\boldsymbol{UB})^{-1}\boldsymbol{BVA}^{-1}$. For the particular case $\boldsymbol{B} = 1$, $\boldsymbol{U} = \mathbf{u}$, and $\boldsymbol{V} = \mathbf{v}^\top$, we have $(\boldsymbol{A} + \mathbf{uv}^\top)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\mathbf{uv}^\top\boldsymbol{A}^{-1}(1 + \mathbf{v}^\top\boldsymbol{A}^{-1}\mathbf{u})^{-1}$. Furthermore
$$\mathbf{x}^\top(\boldsymbol{A} + \mathbf{xx}^\top)^{-1}\mathbf{x} = \frac{\mathbf{x}^\top\boldsymbol{A}^{-1}\mathbf{x}}{1 + \mathbf{x}^\top\boldsymbol{A}^{-1}\mathbf{x}}.$$

[1] 1-VIP-misc-pocket-HL.tex

4.4 Positive (semi)definite Matrices

1. Denote $\boldsymbol{A}_n = (a_{ij})_{nn}$. If $\boldsymbol{A}_n$ is a positive matrix, $|\boldsymbol{A}_n| \le a_{nn}|\boldsymbol{A}_{n-1}|$, and so $|\boldsymbol{A}_n| \le \prod_{i=1}^n a_{ii}$.
2. $\boldsymbol{A}$ and $\boldsymbol{B}$ are real symmetric matrices. Then there exists a orthogonal matrix P such that $P^\top \boldsymbol{A}P$ and $P^\top \boldsymbol{B}P$ are both diagonal if and only if $\boldsymbol{AB} = \boldsymbol{BA}$.
3. $\boldsymbol{A}$ is symmetric. Then $\boldsymbol{A}$ is nonnegative if and only if there exists a matrix $\boldsymbol{C} = (c_{ij})_{rn}$ for $r = rank(\boldsymbol{A})$ such that $\boldsymbol{A} = \boldsymbol{C}^\top \boldsymbol{C} \Longleftrightarrow$ there exists $\boldsymbol{B} = (b_{ij})_{mn}$ such that $\boldsymbol{A} = \boldsymbol{B}^\top \boldsymbol{B} \Longleftrightarrow$ all of eigenvalues of $\boldsymbol{A}$ are nonnegative.
4. $\boldsymbol{A}$ is positive and $\boldsymbol{B}$ is non-negative. Then $|\boldsymbol{A}+\boldsymbol{B}| \ge |\boldsymbol{A}|$ and equality holds if and only if $\boldsymbol{B} = 0$.
5. If $\boldsymbol{A}$ is positive, then
$$f(Y) = \begin{bmatrix} \boldsymbol{A} & Y \\ Y^\top & 0 \end{bmatrix}$$
is negative, where $Y = (y_1, \cdots, y_n)^\top$.
Note that $f(Y) = Y^\top\{(-1)^{2n+2}\boldsymbol{A}^*\}Y$, where $\boldsymbol{A}^* = |\boldsymbol{A}|\boldsymbol{A}^{-1}$.
6. **Cholesky Decomposition**.
For positive matrix $\boldsymbol{A}_{n\times n}$, there exists Cholesky-decomposition $\boldsymbol{A} = \boldsymbol{T}^\top \boldsymbol{T}$, where $\boldsymbol{T}$ is an upper triangular. Also $\boldsymbol{T}$ is unique.
7. Let $X^\top = (x_1, \cdots, x_n)$, where the x_i are n independent $d-$dimensional vectors of random variables, and let $\boldsymbol{A}$ be a positive semidefinite $n\times n$ matrix of rank $r(\ge d)$. Suppose that for each x_i and all $\mathbf{b}(\ne 0)$ and c, prob$[\mathbf{b}^\top x_i = c] = 0$. Then prob$(X^\top \boldsymbol{A}X > 0) = 1$.
8. If $\boldsymbol{A}$ is positive and $\boldsymbol{B}$ is symmetric. Then there exists a nonsingular matrix C such that $C^\top \boldsymbol{A}C =$identity matrix and $C^\top \boldsymbol{B}C = \Lambda$, a diagonal matrix diag$(\lambda_1, \cdots, \lambda_r, 0, \cdots, 0)$, where λ_i is the eigenvalue of $\boldsymbol{A}$.
Proof. There exists a $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{D}^\top \boldsymbol{D}$. Note that $\boldsymbol{D}^{-1\top}\boldsymbol{BD}^{-1}$ is still symmetric. There is a $\boldsymbol{C}$ such that $\boldsymbol{C}^\top \boldsymbol{D}^{-1\top}\boldsymbol{BD}^{-1}\boldsymbol{C} = \Lambda$. Taking $C = \boldsymbol{D}^{-1}\boldsymbol{C}$, the proof follows.
9. Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive. $\boldsymbol{A}-\boldsymbol{B}$ is positive if and only if $\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}$ is positive.
Proof. Denote $\boldsymbol{S} = \boldsymbol{A} - \boldsymbol{B}$. If $\boldsymbol{S}$ is positive, there exists a nonsingular matrix C such that $C^\top \boldsymbol{SC} = \boldsymbol{I}$ and $C^\top \boldsymbol{BC} = \Lambda$. It follows that $\boldsymbol{A} = C^{-1\top}(\Lambda + \boldsymbol{I})C^{-1}$ and $\boldsymbol{B} = C^{-1\top}\Lambda C^{-1}$, and then $\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1} = C^{-1\top}\{\Lambda^{-1} - (\Lambda+\boldsymbol{I})^{-1}\}C^{-1} \ge \boldsymbol{O}$. Conversely, the proof for significant condition becomes trivial because the above arguments.
The conclusion can be generalized to **nonnegative** *case since we can consider* $\boldsymbol{A} - \boldsymbol{B} + 1/n\boldsymbol{I}$ *and finally let* $n \to \infty$. *Without the assumption of* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *being positive, the conclusion is false, see for example,* $\boldsymbol{A} = \boldsymbol{I}$ *and* $\boldsymbol{B} = -0.5\boldsymbol{I}$.

4.5 Idempotent Matrices

A matrix $\boldsymbol{A}$ is idempotent if $\boldsymbol{A}^2 = \boldsymbol{A}$. A symmetric idempotent matrix is called a projection matrix.

1. If $\boldsymbol{A}$ is a projection matrix of rank r, then it can be expressed in the form
$$\boldsymbol{A} = \sum_{i=1}^r \mathbf{t}_i \mathbf{t}_i^\top.$$
where $\mathbf{t}_1, \cdots, \mathbf{t}_r$ form an orthonormal set.
2. If $\boldsymbol{A}$ is a projection matrix, then $r(\boldsymbol{A}) = \text{tr}(\boldsymbol{A})$.
3. If $\boldsymbol{A}$ is idempotent, then so is $I - \boldsymbol{A}$.

4.6 Vector and Matrix Differentiation

$$\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \qquad \frac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$$
$$\frac{\partial \mathbf{x}^\top \boldsymbol{A}\mathbf{x}}{\partial \mathbf{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\mathbf{x} \qquad \frac{\partial \mathbf{x}^\top \boldsymbol{A}\mathbf{y}}{\partial \mathbf{x}} = \boldsymbol{A}\mathbf{y}$$
$$\frac{\partial |\boldsymbol{X}|}{\partial x_{ij}} = \begin{cases} X_{ij} & \text{if all elements of } \boldsymbol{X} \text{ are distinct} \\ \left.\begin{array}{ll} X_{ii} & \text{if } i = j \\ 2X_{ij} & \text{otherwise} \end{array}\right\} & \boldsymbol{X} \text{ is symmetric} \end{cases}$$
$$\frac{\partial tr\boldsymbol{XY}}{\partial \boldsymbol{X}} = \begin{cases} \boldsymbol{Y}^\top & \text{if all elements of } \boldsymbol{X} \text{ are distinct} \\ \boldsymbol{Y} + \boldsymbol{Y}^\top - Diag(\boldsymbol{Y}) & \text{if } \boldsymbol{X} \text{ is symmetric} \end{cases}$$
$$\frac{\partial \boldsymbol{X}^{-1}}{\partial x_{ij}} = \begin{cases} \boldsymbol{X}^{-1}\boldsymbol{J}_{ij}\boldsymbol{X}^{-1} & \text{if all elements of } \boldsymbol{X} \text{ are distinct} \\ \left.\begin{array}{ll} \boldsymbol{X}^{-1}\boldsymbol{J}_{ii}\boldsymbol{X}^{-1} & \text{if } i = j \\ \boldsymbol{X}^{-1}(\boldsymbol{J}_{ij} + \boldsymbol{J}_{ji})\boldsymbol{X}^{-1} & \text{otherwise} \end{array}\right\} & \boldsymbol{X} \text{ is symmetric} \end{cases}$$

where X_{ij} denotes the cofactor of x_{ij} in $\boldsymbol{X}$ and $\boldsymbol{J}_{ij}$ denotes a matrix with 1 in the (i,j)th place and zeros elsewhere.

4.7 Basic Concepts and Facts

1. For $s = m$ and $m \ge n$, $|\lambda E - \boldsymbol{AB}| = \lambda^{m-n}|\lambda E - \boldsymbol{BA}|$.
2. $\boldsymbol{A}$ is $n-$order matrix. If $|a_{ii}| \ge \sum_{i\ne j}|a_{ij}|$ for $i = 1, \cdots, n$, then $|\boldsymbol{A}| \ne 0$; If $a_{ii} \ge \sum_{i\ne j}|a_{ij}|$ for $i = 1, \cdots, n$, then $|\boldsymbol{A}| > 0$.
3. $\boldsymbol{A}$ is $n-$order matrix. There exists a orthogonal matrix P such that $P^\top \boldsymbol{A}P$ is trigonal matrix if and only if all eigenvalues of $\boldsymbol{A}$ are real.
4. If $\boldsymbol{A}$ is orthogonal and its eigenvalues are all of real, then $\boldsymbol{A}$ must be symmetric and $\boldsymbol{A}^2 = E$.
5. Suppose that $\{\boldsymbol{A}_k\}$ is a sequence of real symmetric matrices, and $\boldsymbol{A}_i\boldsymbol{A}_j = \boldsymbol{A}_j\boldsymbol{A}_i$ for $i \ne j$. Then there exists a orthogonal matrix P such that $P^\top \boldsymbol{A}_k P$ are all diagonal.
6. Any real inverable matrix $\boldsymbol{A}$ can be decomposed into SS_1S_2, where S is positive, S_1 and S_2 are real symmetric orthogonal.

1. A square matrix $\boldsymbol{A}$ such that $\boldsymbol{A}^k = 0$ for some integer k is called $\boxed{\textit{nilpotent}}$ matrix, and the smallest positive integral exponent k such that $\boldsymbol{A}^k = 0$ is called the *index* of $\boldsymbol{A}$.
2. $\boldsymbol{A} = -\boldsymbol{A}^\top$ is called $\boxed{\textit{skew-symmetric}}$.
3. A matrix $\boldsymbol{A}$ with complex elements is said to be $\boxed{\textit{Hamitian}}$ if $\boldsymbol{A} = \boldsymbol{A}^*$, and *skew-Hermitian* if $\boldsymbol{A} = -\boldsymbol{A}^*$.
4. $\boldsymbol{A}$ is $\boxed{\textit{equivalent}}$ to $\boldsymbol{B}$ if $\boldsymbol{B}$ can be obtained from $\boldsymbol{A}$ by the successive application of finitely many elementary row and column operations, and we write $\boldsymbol{A} \overset{\text{E}}{=} \boldsymbol{B}$.
5. $\boldsymbol{A} \overset{\text{E}}{=} \boldsymbol{B}$ iff $\boldsymbol{A} = S\boldsymbol{A}T$, where S and T are $m\times m$ and $n\times n$ non-singular matrices, respectively.
6. Every nonzero $\boldsymbol{A}$ is *equivalent* to $\begin{bmatrix} I_r & O_{r,n-r} \\ O_{m-r,r} & O_{m-r,m-r} \end{bmatrix}$.
7. Matrices of the form $\begin{bmatrix} I_r & O \\ O & O \end{bmatrix}$ are called $\boxed{\textit{canonical matrices}}$.
8. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are two square matrices of order n and m, respectively, the matrix
$$\boldsymbol{A} \dot{+} \boldsymbol{B} = \begin{bmatrix} \boldsymbol{A} & O \\ O & \boldsymbol{B} \end{bmatrix}$$
is called their *direct sum*.
9. $\boldsymbol{A}_{ij}$ is called the $\boxed{\textit{cofactor}}$ of element a_{ij}.
10. (The Hamilton-Cayley Theorem) Let $\boldsymbol{A}$ be an $n\times n$ matrix and let
$$f(\lambda) \equiv (-1)^n\{\lambda^n - p_1\lambda^{n-1} + \cdots + (-1)^n p_n\}$$
be the characteristic function of $\boldsymbol{A}$. Then
$$\boldsymbol{A}^n - p_1\boldsymbol{A}^{n-1} + \cdots + (-1)^n p_n I_n = O_n.$$
11. Denote I_n by $\boldsymbol{A}_0$, one successively computes $c_1, \boldsymbol{A}_1, c_2, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_{n-1}, c_n$ by the two formulas
$$c_k = (1/k)\text{tr}(\boldsymbol{A}\boldsymbol{A}_k) \quad \boldsymbol{A}_k = \boldsymbol{A}\boldsymbol{A}_{k-1} - c_k I$$
Then $\boldsymbol{A}^{-1} = \boldsymbol{A}_{n-1}/c_n$.
12. If $\lambda_1, \lambda_2, \cdots, \lambda_n$ are the characteristic roots, distinct or not, of an $n\times n$ matrix $\boldsymbol{A}$, and if $g(\boldsymbol{A})$ is any polynomial function of $\boldsymbol{A}$, then the characteristic roots of $g(\boldsymbol{A})$ are $g(\lambda_1), \cdots, g(\lambda_n)$.
13. $\boldsymbol{A}$ is $\boxed{\textit{similar}}$ to $\boldsymbol{B}$ if and only if there exists a nonsingular matrix P such that $\boldsymbol{A} = P^{-1}\boldsymbol{B}P$. We write $\boldsymbol{A} \overset{\text{S}}{=} \boldsymbol{B}$.
14. Any square matrix $\boldsymbol{A}$ is *similar* to an upper triangular matrix whose diagonal elements are the eigenvalues of $\boldsymbol{A}$.
15. A square matrix $\boldsymbol{B}$ is said to be $\boxed{\textit{congruent}}$ to $\boldsymbol{A}$ if and only if there exists a nonsingular matrix P such that $\boldsymbol{A} = P^\top \boldsymbol{B}P$. We write $\boldsymbol{A} \overset{\text{C}}{=} \boldsymbol{B}$.
16. An elementary row operation applied to a square matrix, and followed by the corresponding elementary column operation, is called an *elementary cogredient operation* on the matrix.
17. Every symmetric $\boldsymbol{R}$ matrix $\boldsymbol{A}$ of rank r is *congruent* to a matrix of the form diag(I_r, O).
18. Any $n\times n$ real symmetric matrix $\boldsymbol{A}$ is orthogonal similar to a diagonal matrix whose diagonal elements are the eigenvalues of $\boldsymbol{A}$.
19. | | |
|---|---|
| $\boldsymbol{AB} = \boldsymbol{BA}$ | commutative |
| $\boldsymbol{AB} = -\boldsymbol{BA}$ | anti-commute |
| $\boldsymbol{A}^2 = I$ | involutory |

4.8 Optimization and Inequalities

1. Consider the matrix function f, where

$$f(\boldsymbol{X}) = -\log|\boldsymbol{X}| + \mathrm{tr}(\boldsymbol{X}^{-1}\boldsymbol{A}).$$

If $\boldsymbol{A} > \boldsymbol{O}$, then, subject to $\boldsymbol{X} > \boldsymbol{O}$, $f(\boldsymbol{X})$ is minimized uniquely at $\boldsymbol{X} = \boldsymbol{A}$.

2. Let $f : \theta \longrightarrow f(\theta)$ be a real-valued function with domain Θ, and let $g : \theta \longrightarrow g(\theta) = \phi$ be a bijective (one-to-one) function from Θ onto Φ. Since g is bijective, it has an inverse, g^{-1}, say, and we can define $h(\phi) = f(g^{-1}(\theta))$ for $\phi \in \Phi$.

   (a) If $f(\theta)$ attains a maximum at $\theta = \widehat{\theta}$, $h(\phi)$ attains its maximum at $\widehat{\phi} = g(\theta)$.

   (b) If the maximum of $f(\theta)$ occurs uniquely at $\widehat{\theta}$, then the maximum of $h(\phi)$ occurs uniquely at $\widehat{\phi}$.

3. *Frobenius norm approximation*. Let $\boldsymbol{B}$ be a $p \times q$ matrix of rank r with singular value decomposition $\sum_{i=1}^{r} \delta_i l_i m_i^\top$, and let $\boldsymbol{C}$ be a $p \times q$ matrix of ranks $s (s < r)$. Then

$$\|\boldsymbol{B} - \boldsymbol{C}\|^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}(b_{ij} - c_{ij})^2$$

is minimized when

$$\boldsymbol{C} = \boldsymbol{B}_{(s)} = \sum_{i=1}^{s} \delta_i l_i m_i^\top.$$

The minimum value is $\sum_{i=s+1}^{r} \delta_i^2$.

4. Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$, and a corresponding set of orthogonal eigenvectors $\mathbf{t}_1, \cdots, \mathbf{t}_n$. Define $T_k = (\mathbf{t}_1, \cdots, \mathbf{t}_k)$ $(k = 1, \cdots, n-1)$ and $T = (\mathbf{t}_1, \cdots, \mathbf{t}_n)$. Then, if we assume that $\mathbf{x} \neq 0$, we have the following:

   (a)

$$\sup_{\mathbf{x}} \frac{\mathbf{x}^\top \boldsymbol{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_1,$$

   and the supremum is attained if $\mathbf{x} = \mathbf{t}_1$.

   (b)

$$\sup_{T_k^\top \mathbf{x}=0} \frac{\mathbf{x}^\top \boldsymbol{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_{k+1},$$

   and the supremum is attained if $\mathbf{x} = \mathbf{t}_{k+1}$.

   (c)

$$\inf_{\mathbf{x}} \frac{\mathbf{x}^\top \boldsymbol{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_n,$$

   and the infimum is attained if $\mathbf{x} = \mathbf{t}_n$.

   (d) If $T_{n-k} = (\mathbf{t}_{n-k+1}, \cdots, \mathbf{t}_n)$

$$\inf_{T_{n-k}^\top \mathbf{x}=0} \frac{\mathbf{x}^\top \boldsymbol{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_{n-k},$$

   and the infimum is attained if $\mathbf{x} = \mathbf{t}_{n-k}$.

   (e) *Courant-Fischer min-max theorem.*

$$\inf_{L_{n\times k}} \sup_{L^\top \mathbf{x}=0} \frac{\mathbf{x}^\top \boldsymbol{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_{k+1},$$

   and the result is attained if $L = T_k$ and $\mathbf{x} = \mathbf{t}_{k+1}$.

   (f)

$$\sup_{L_{n\times k}} \inf_{L^\top \mathbf{x}=0} \frac{\mathbf{x}^\top \boldsymbol{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_{n-k},$$

   and the result is attained if $L = T_{n-k}$ in (d) and $\mathbf{x} = \mathbf{t}_{n-k}$.

5. Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix and let $\boldsymbol{D}$ be any $n \times n$ positive definite matrix. Let $\gamma_1 \geq \cdots \geq \gamma_n$ be eigenvalues of $\boldsymbol{D}^{-1}\boldsymbol{A}$ with corresponding eigenvectors $\mathbf{v}_1, \cdots, \mathbf{v}_n$. Then

$$\sup_{\mathbf{x}} \frac{\mathbf{x}^\top \boldsymbol{A}\mathbf{x}}{\mathbf{x}^\top \boldsymbol{D}\mathbf{x}} = \gamma_1, \quad \text{and } \inf_{\mathbf{x}} \frac{\mathbf{x}^\top \boldsymbol{A}\mathbf{x}}{\mathbf{x}^\top \boldsymbol{D}\mathbf{x}} = \gamma_n,$$

with the bounds being attained when $\mathbf{x} = \mathbf{v}_1$ and $\mathbf{x} = \mathbf{v}_n$, respectively.

6. If $\boldsymbol{D}$ is positive definite, then for any $\mathbf{a}$

$$\sup_{\mathbf{x}} \frac{(\mathbf{a}^\top \mathbf{x})^2}{\mathbf{x}^\top \boldsymbol{D}\mathbf{x}} = \mathbf{a}^\top \boldsymbol{D}^{-1}\mathbf{a}.$$

The supremum occurs when $\mathbf{x}$ is proportional to $\boldsymbol{D}^{-1}\mathbf{a}$.

7. Let $\boldsymbol{M}$ and $\boldsymbol{N}$ be positive definite, then

$$\sup_{\mathbf{x},\mathbf{y}} \frac{\mathbf{x}^\top \boldsymbol{L}\mathbf{x}}{\mathbf{x}^\top \boldsymbol{M}\mathbf{x} \cdot \mathbf{y}^\top \boldsymbol{N}\mathbf{y}} = \theta_{max},$$

where θ_{max} is the largest eigenvalue of $\boldsymbol{M}^{-1}\boldsymbol{L}^\top \boldsymbol{N}^{-1}\boldsymbol{L}$. The supremum occurs when $\mathbf{x}$ is an eigenvector of $\boldsymbol{M}^{-1}\boldsymbol{L}^\top \boldsymbol{N}^{-1}\boldsymbol{L}$ corresponding to θ_{max}, and $\mathbf{y}$ is an eigenvector of $\boldsymbol{M}^{-1}\boldsymbol{L}^\top \boldsymbol{M}^{-1}\boldsymbol{L}$ corresponding to θ_{max}.

8. Let $\boldsymbol{C}$ be $p \times q$ matrix of rank m and let $\rho_1^2 \geq \cdots \geq \rho_m^2 > 0$ be the nonzero eigenvalues of $\boldsymbol{C}\boldsymbol{C}^\top$. Let $\mathbf{t}_1, \cdots, \mathbf{t}_m$ be the corresponding eigenvectors of $\boldsymbol{C}\boldsymbol{C}^\top$ and let $\mathbf{w}_1, \cdots, \mathbf{w}_m$ be the corresponding eigenvectors of $\boldsymbol{C}^\top \boldsymbol{C}$. If $T_k = (\mathbf{t}_1, \cdots, \mathbf{t}_k)$ and $W_k = (\mathbf{w}_1, \cdots, \mathbf{w}_k)$ $(k < m)$, then

$$\sup_{\mathbf{T}_\mathbf{k}^\top \mathbf{x}=\mathbf{0}, \mathbf{W}_\mathbf{k}^\top \mathbf{y}=\mathbf{0}} \frac{(\mathbf{x}^\top \boldsymbol{C}\mathbf{y})^2}{\mathbf{x}^\top \mathbf{x} \cdot \mathbf{y}^\top \mathbf{y}} = \rho_{k+1}^2,$$

and the supremum occurs when $\mathbf{x} = \mathbf{t}_{k+1}$ and $\mathbf{y} = \mathbf{w}_{k+1}$.

9. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be an $n \times n$ symmetric matrices with eigenvalues $\rho_1(\boldsymbol{A}) \geq \cdots \geq \rho_n(\boldsymbol{A})$ and $\rho_1(\boldsymbol{B}) \geq \cdots \geq \rho_n(\boldsymbol{B})$, respectively. If $\boldsymbol{A} - \boldsymbol{B} \geq \boldsymbol{O}$, then we have the following:

   (a) $\rho_i(\boldsymbol{A}) \geq \rho_i(\boldsymbol{B})$ $(i = 1, \cdots, n)$

   (b) $\mathrm{tr}(\boldsymbol{A}) \geq \mathrm{tr}(\boldsymbol{B})$

   (c) $|\boldsymbol{A}| \geq |\boldsymbol{B}|$

   (d) $\|\boldsymbol{A}\| \geq \|\boldsymbol{B}\|$, where $\|\boldsymbol{A}\| = \{\mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^\top)\}^{1/2}$.

10. Let $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{A} - \boldsymbol{B}$ be an $n \times n$ positive semidefinite matrices, with $r(\boldsymbol{B}) \leq r$, and let $\rho_i(\cdot)$ represent the ith largest eigenvalues. Then

$$\rho_i(\boldsymbol{A} - \boldsymbol{B}) \geq \begin{cases} \rho_{r+i}(\boldsymbol{A}) & i = 1, \cdots, n-r \\ 0 & i = n-r+1, \cdots, n \end{cases}$$

Equality occurs if

$$\boldsymbol{B} = \boldsymbol{B}_0 = \sum_{i=1}^{r} \rho_i(\boldsymbol{A})\mathbf{t}_i\mathbf{t}_i^\top,$$

where $\mathbf{t}_1, \cdots, \mathbf{t}_n$ are orthogonal eigenvectors corresponding to $\rho_1(\boldsymbol{A}), \cdots, \rho_n(\boldsymbol{A})$.

4.9 Jacobians and Transformations

1. If the distinct elements of a symmetric $d \times d$ matrix $\boldsymbol{A}$ have a joint density function of the form $g(\lambda_1, \cdots, \lambda_d)$ where $\lambda_1 \geq \cdots \geq \lambda_d$ are the eigenvalues of $\boldsymbol{A}$, then the joint density function of the eigenvalues is

$$\pi^{d^2/2} g(\lambda_1, \cdots, \lambda_d) \left\{ \prod_{j<k} (\lambda_j - \lambda_k) \right\} \Big/ \Gamma_d(d/2)$$

where $\Gamma_d(d/2) = \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma(\frac{1}{2}[d+1-j])$.

2. Let $\boldsymbol{X}$ be an $m \times n$ matrix of distinct random variables and let $\boldsymbol{Z} = a(\boldsymbol{X})$, where $\boldsymbol{Z}$ is $m \times n$ and a is a bijective function. Then there exists an inverse function $b = a^{-1}$, so that $\boldsymbol{X} = b(\boldsymbol{Z})$. If $\boldsymbol{X}$ has density f and $\boldsymbol{Z}$ has density density g, then

$$g(\boldsymbol{Z}) = f(b[\boldsymbol{Z}]) \left| \frac{d\boldsymbol{X}}{d\boldsymbol{Z}} \right|,$$

where $d\boldsymbol{X}/d\boldsymbol{Z}$ represents the Jacobian of the transformation from $\boldsymbol{X}$ to $\boldsymbol{Z}$.

   (a) If $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z}\boldsymbol{B}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are $m \times m$ and $n \times n$ nonsingular matrices, respectively, then

$$\frac{d\boldsymbol{X}}{d\boldsymbol{Z}} = |\boldsymbol{A}|^n |\boldsymbol{B}|^m.$$

   (b) If $\boldsymbol{X}$ and $\boldsymbol{Z}$ are $n \times n$ symmetric matrices, $\boldsymbol{A}$ is nonsingular matrices, and $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z}\boldsymbol{A}^\top$

$$\frac{d\boldsymbol{X}}{d\boldsymbol{Z}} = |\boldsymbol{A}|^{n+1}$$

   (c) Let $\boldsymbol{E}$ and $\boldsymbol{H}$ be $d \times d$ positive definite matrices, and let $\boldsymbol{Z} = \boldsymbol{E} + \boldsymbol{H}$ and $\boldsymbol{V} = (\boldsymbol{E} + \boldsymbol{H})^{-1/2}\boldsymbol{H}(\boldsymbol{E} + \boldsymbol{H})^{-1/2}$.

$$\frac{d(\boldsymbol{H}, \boldsymbol{E})}{d(\boldsymbol{V}, \boldsymbol{Z})} = |\boldsymbol{Z}|^{(d+1)/2}.$$

4.10 Generalized Inverse

4.10.1 $g-$inverse

If $\boldsymbol{A}\boldsymbol{G}\boldsymbol{A} = \boldsymbol{A}$, $\boldsymbol{G}$ is called a generalized inverse ($g-$inverse).

1. For each matrix $\boldsymbol{A} = \boldsymbol{A}_L \boldsymbol{A}_R \triangleq \boldsymbol{B}\boldsymbol{C}, \boldsymbol{A}^- = \boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{C}^\top)^{-1}(\boldsymbol{B}^\top\boldsymbol{B})^{-1}\boldsymbol{B}^\top$.
2. $r(\boldsymbol{A}) = r(\boldsymbol{A}^-) = r(\boldsymbol{A}\boldsymbol{A}^-) = r(\boldsymbol{A}^-\boldsymbol{A})$.
3. If $\boldsymbol{A}$ is an $m \times n$ matrix of rank m, then $\boldsymbol{A}^- = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}$ (right-inverse) and $\boldsymbol{A}\boldsymbol{A}^- = I_m$. If $r(\boldsymbol{A}) = n$, then $\boldsymbol{A}^- = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top$ (left-inverse) and $\boldsymbol{A}^-\boldsymbol{A} = I_n$.
4. It is not always true that $(\boldsymbol{G}H)^- = \boldsymbol{H}^-\boldsymbol{G}^-$ for all matrices $\boldsymbol{H}, G$.
5. Let $\boldsymbol{B}$ be an $m \times r$ matrix of rank r and $\boldsymbol{C}$ be an $r \times m$ matrix of rank r; then $(\boldsymbol{B}\boldsymbol{C})^- = \boldsymbol{C}^-\boldsymbol{B}^-$.
6. $(\boldsymbol{A}^\top\boldsymbol{A})^- = \boldsymbol{A}^-(\boldsymbol{A}^\top)^-$ for any matrix $\boldsymbol{A}$.
7. Let $\boldsymbol{P}$ be an $m \times m$ orthogonal matrix, $\boldsymbol{Q}$ be an $n \times n$ orthogonal matrix, and $\boldsymbol{A}$ is any $m \times n$ matrix. Then $(\boldsymbol{P}\boldsymbol{A}\boldsymbol{Q})^- = \boldsymbol{Q}^-\boldsymbol{A}^-\boldsymbol{P}^-$.

4.10.2 $c-$inverse

1. $r(\boldsymbol{X}^c) \geq r(\boldsymbol{X}) = r(\boldsymbol{X}\boldsymbol{X}^c) = r(\boldsymbol{X}^c\boldsymbol{X})$ for any matrix $\boldsymbol{X}$.
2. $\boldsymbol{X}^c\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}^c$ are idempotent matrices.
3. If $\boldsymbol{X}^c$ is any $c-$inverse of $\boldsymbol{X}$, then $(\boldsymbol{X}^c)^\top$ is a $c-$inverse of $\boldsymbol{X}^\top$.
4. For any $m \times n$ matrix $\boldsymbol{X}$ of rank $r > 0$, define
$$\boldsymbol{K} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^c\boldsymbol{X}^\top.$$
Then $\boldsymbol{K}$ is invariant for any $c-$inverse of $\boldsymbol{X}^\top\boldsymbol{X}$.
5. $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^c\boldsymbol{X}^\top = \boldsymbol{X}\boldsymbol{X}^-$ for any $c-$inverse $(\boldsymbol{X}^\top\boldsymbol{X})^c$ of $\boldsymbol{X}^\top\boldsymbol{X}$.
6. $r(\boldsymbol{K}) = r(\boldsymbol{X})$.
7. $\boldsymbol{K}\boldsymbol{X} = \boldsymbol{X}; \boldsymbol{X}^\top\boldsymbol{K} = \boldsymbol{X}^\top$
8. $(\boldsymbol{X}^\top\boldsymbol{X})^c\boldsymbol{X}^\top$ is a $c-$inverse of $\boldsymbol{X}$ for any $c-$inverse of $\boldsymbol{X}^\top\boldsymbol{X}$.
9. $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^c$ is a $c-$inverse of $\boldsymbol{X}^\top$ for any $c-$inverse of $\boldsymbol{X}^\top\boldsymbol{X}$.

4.11 Linear Equations

Let $\boldsymbol{A}$ be an $m \times n$ matrix and $\boldsymbol{A}^c$ be any $c-$inverse of $\boldsymbol{A}$. Suppose a solution exists to the system $\boldsymbol{A}\mathbf{x} = g$. For each $n \times 1$ vector $\mathbf{h}$, the vector $\mathbf{x}_0$ is a solution, where
$$\mathbf{x}_0 = \boldsymbol{A}^c g + (I_n - \boldsymbol{A}^c\boldsymbol{A})\mathbf{h}. \tag{1}$$
Also, every solution to the system can be written in the form of Equation (1) for some $n \times 1$ vector $\mathbf{h}$.

1. If $\boldsymbol{A}$ is an $m \times m$ symmetric matrix such that $\mathbf{1}^\top\boldsymbol{A} = \mathbf{0}$, then
$$\begin{bmatrix} \boldsymbol{A} \\ \mathbf{1}^\top \end{bmatrix}^- = \left[\boldsymbol{A}^-, \frac{1}{m}\mathbf{1}\right]$$
2. If $\boldsymbol{A}$ is an $m \times m$ symmetric matrix of rank $m-1$ such that $\mathbf{1}^\top\boldsymbol{A} = \mathbf{0}$, then $\boldsymbol{B} = \boldsymbol{A} + \mathbf{1}\mathbf{1}^\top/n$ is nonsingular and its inverse is $\boldsymbol{A}^- + \boldsymbol{J}/n$; Meanwhile
$$\begin{bmatrix} \boldsymbol{A} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix}^- = \begin{bmatrix} \boldsymbol{A}^- & \frac{1}{n}\mathbf{1} \\ \frac{1}{n}\mathbf{1}^\top & 0 \end{bmatrix}.$$

5 Analysis

analysis.tex

1. $f(x)$ is bound and $g(x)$ is differentiable on $[a,b]$. $g(\lambda) = 0$ for some $\lambda \neq 0$, if $|g(x)f(x) + \lambda g'(x)| \leq |g(x)|$, then $g(x) = 0$ for all $x \in [a,b]$.
2. $f(x)$ is monotone on $[0,\infty]$ and $\int_0^\infty f(x)dx$ is well-defined. Then
$$\lim_{h\to 0^+} h\sum_{n=1}^{\infty} f(nh) = \int_0^\infty f(x)dx.$$
3. $f(x)$ is $2n$-differentiable on $[a,b]$, and $|f^{(2n)}(x)| \leq M$, $f^{(m)}(a) = f^{(m)}(b) = 0$ for $m = 0,\cdots,n-1$. Then
$$|\int_a^b f(x)dx| \leq \frac{(n!)^2 M}{(2n)!(2n+1)!}(b-a)^{2n+1}.$$
4. $f(x)$ and $g(x)$ are bound on any sub-interval of $[0,+\infty)$, and satisfy that $g(x+T) > g(x)$ for some $T > 0$ and any $x > 0$, $g(x) \to +\infty$. In addition,
$$\lim_{x\to\infty}\frac{f(x+T)-f(x)}{g(x+T)-g(x)} = l,$$
where l may be $+\infty$. Then
$$\lim_{x\to\infty}\frac{f(x)}{g(x)} = l.$$
5. $f(x)$ and $g(x)$ are bound on any sub-interval of $[0,+\infty)$, and satisfy that $0 < g(x+T) < g(x)$ for some $T > 0$ and any $x > 0$, $\lim_{x\to\infty} g(x) = 0$. In addition,
$$\lim_{x\to\infty}\frac{f(x+T)-f(x)}{g(x+T)-g(x)} = l,$$
where l may be $+\infty$. Then
$$\lim_{x\to\infty}\frac{f(x)}{g(x)} = l.$$
6. $f'(x)$ is absolutely continuous on $[a,b]$, then for any $c \in (a,b)$ and $p > 1$
$$\int_a^b |f''(x)|^p dx \geq \left\{\frac{p-1}{2p-1}(b-a)\right\}^{1-p}\left|\frac{f(b)-f(c)}{b-c} - \frac{f(c)-f(a)}{c-a}\right|^p.$$
7. Positive series $\sum u_n$
   - comparing principle;
   - integal decision;
   - $\lim u_n/u_{n+1} = \rho$, $\rho > 1$ converges and $\rho < 1$ deverges;
   - Cauchy criterion: $\limsup u_n^{1/n} = \rho$, $\rho > 1$ converges and $\rho < 1$ deverges;
   - $u_n/u_{n+1} = \lambda + \mu/n + o(\theta_n/n^{1+t})$, $\lambda > 1$ converges, $\lambda < 1$ deverges, $\lambda = 1$ $\mu > 1$ converges and $\mu < 1$ deverges;
   - $\sum |b_n| < \infty$, $u_n/u_{n+1} = \lambda + 1/n + o(b_n)$, deverges;
   - $u_{n+1}/u_n = 1 - \alpha/n + O(1/n^\lambda)$, $\alpha > 1$ converges, $\alpha < 1$ deverges;
   - $u_{n+1}/u_n = 1 - 1/n - \alpha_n'/n\log n$, $\alpha_n' \geq \alpha > 1$ converges, $\alpha_n' \leq \alpha < 1$ deverges.
8. (Roll Theorem) $f'(x)$ is bound on finite or infinite interval (a,b), and $\lim_{x\to a^+} f(x) = \lim_{x\to b^-} f(x)$. Then there exists at least one $c \in (a,b)$ such that $f'(c) = 0$.
9. (Dabu Theorem) $m < |f'(x)| < M$ Then for any $\mu \in (m,M)$, there exists $x_\mu \in (a,b)$ such that $f'(x_\mu) = \mu$.
10. $f(x)$ is differentiable on and $|f'(x)| \leq M$ on (a,b), then
$$\left|\frac{1}{b-a}\int_a^b f(x)dx - \frac{f(b)+f(a)}{2}\right| \leq \frac{M(b-a)}{4}\left\{1 - \left(\frac{f(b)-f(a)}{M(b-a)}\right)^2\right\}.$$
11. From Joseph Edward (1954).
$$\int_0^1 \frac{(-\log x)^p}{(1-x)^2}dx = \int_0^\infty y^p \sum_{n=1}^{\infty} ny^{-ny}dy = p!\sum_{k=1}^{\infty}\frac{1}{k^p}$$
12.
$$\int_0^1 \frac{(-\log x)^p}{(1+x)^2}dx = p!(1-2^{1-p})\sum_{k=1}^{\infty}\frac{1}{k^p}$$

The relationship of some important inequalities

Denote $M_r(a) = (\sum_{i=1}^n p_i a_i^r)^{1/r}$ for $a = (a_1,\cdots,a_n)^\top$ with $a_i \geq 0$ and $r \geq 1$, $\{p_i\}$ are weight ($\sum p_i = 1$). $G(a) = \prod a_i^{p_i}$.
Cauchy$\Longrightarrow M_{2r}(a) > M_r(a) \Longrightarrow M_r(a) \to G(a)$ by letting $r \to 0 \Longrightarrow$ Hölder

$$\sum_{k=1}^n a_k^\alpha b_k^\beta \cdots l_k^\lambda \leq (\sum_{k=1}^n a_k)^\alpha(\sum_{k=1}^n b_k)^\beta\cdots(\sum_{k=1}^n l_k)^\lambda$$
for $\alpha,\beta,\cdots,\lambda > 0$ and $\alpha+\beta+\cdots+\lambda = 1$.

$$\text{Hölder} \Rightarrow \begin{cases} \text{(Minkowski)}\left\{\sum_{j=1}^k a_j^r\right\}^{1/r} \leq \left\{\sum_{j=1}^k a_j^r\right\}^{1/r} & r > 1 \\ \text{(Jensen)}\ (\sum_1^n a_k^s)^{1/s} \leq (\sum_1^n a_k^r)^{1/r} & 0 < r < s \\ \text{(Liap)}\ \{M_s(a)\}^s \leq [\{M_r(a)\}^r]^{\frac{(t-s)}{t-r}}[\{M_t(a)\}^t]^{\frac{(s-r)}{t-r}} & 0 < r < s \\ \text{(Increasing property of)}\ f(x) = M_x(a) & \end{cases}$$

6 Good Examples in Linear Models

linear.tex

1. Consider the linear model

$$\underline{Y} = \boldsymbol{X}\underline{\beta} + \boldsymbol{Z}\underline{\gamma} + \underline{e} \quad \underline{e} \sim N(\bar{0}, \sigma^2\mathbf{I}).$$

where $\underline{Y}$ is $n\times 1$, $\underline{\gamma}$ $q\times 1$, $\underline{\beta}$ is $p\times 1$, $\boldsymbol{X}$ is $n\times p$, $\boldsymbol{Z}$ is $n\times q$, $[\boldsymbol{X},\boldsymbol{Z}]$ is of rank $p+q$ and $n > p+q$.

   (a) Show that $\mathbf{Z}^\top(\mathbf{I}-\boldsymbol{X}\boldsymbol{X}^-)\boldsymbol{Z}$ is positive
   $r[\mathbf{Z}^\top(\mathbf{I}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top] = r\{(\boldsymbol{X},\boldsymbol{Z})^\top[\mathbf{I}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top]\} \geq r\{(\boldsymbol{X},\boldsymbol{Z})^\top\} + r\{\mathbf{I}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\} - n = p+q+(n-p)-n = q$ $\mathbf{Z}^\top[\mathbf{I}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top]\mathbf{Z}$ is full rank and nonnegative definite.

   (b) Show that the MLE of $\underline{\beta}$ and $\underline{\gamma}$ are:

$$\widehat{\underline{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\underline{Y}-\boldsymbol{Z}\widehat{\underline{\gamma}}) \text{ and } \widehat{\underline{\gamma}} = \{\mathbf{Z}^\top(\mathbf{I}-\boldsymbol{X}\boldsymbol{X}^-)\mathbf{Z}\}^{-1}\mathbf{Z}^\top(\mathbf{I}-\boldsymbol{X}\boldsymbol{X}^-)\underline{Y}$$

   (c) How would the estimators of $\underline{\beta}$ and $\underline{\gamma}$ change if $\boldsymbol{X}$ and $\boldsymbol{Z}$ were orthogonal? Find the joint distribution of the estimators.

2. Consider the linear model

$$\begin{bmatrix} \underline{Y}_1 \\ \underline{Y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & 0 \\ 0 & \boldsymbol{X}_2 \end{bmatrix}\begin{bmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \end{bmatrix} + \underline{e} \quad \underline{e} \sim N(\bar{0}, \sigma^2\mathbf{I}).$$

where $\underline{Y}_i$ is $n_i\times 1$, $\underline{\beta}_i$ is $p\times 1$, $\boldsymbol{X}_i$ is $n_i \times p$ of rank p and $n_1+n_2 = n$. Now, consider the following three estimators

$$\widehat{\underline{\beta}}_1 = (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\underline{Y}_1,\ \widehat{\underline{\beta}}_2 = (\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\underline{Y}_2,\ \&\widehat{\underline{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\underline{Y}$$

We wish to test $H_0 : \underline{\beta}_1 = \underline{\beta}_2$ vs $H_a : \underline{\beta}_1 \neq \underline{\beta}_2$. Two possible test statistics are

$$F_1 = \frac{(\widehat{\underline{\beta}}_1 - \widehat{\underline{\beta}}_2)^\top\{(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\}^{-1}(\widehat{\underline{\beta}}_1 - \widehat{\underline{\beta}}_2)}{SSE_1 + SSE_2} \times \frac{n-2p}{p}$$

and

$$F_2 = \frac{(\widehat{\underline{\beta}} - \widehat{\underline{\beta}}_1)^\top\{(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\}^{-1}(\widehat{\underline{\beta}} - \widehat{\underline{\beta}}_1)}{SSE_1 + SSE_2} \times \frac{n-2p}{p}$$

   (a) Show that

$$\begin{aligned}&(\boldsymbol{X}_1^\top\boldsymbol{X}_1)\{(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\}(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1} \\ &= (\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\{(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\}^{-1}(\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}(\boldsymbol{X}^\top\boldsymbol{X}).\end{aligned}$$

   (b) Show that $F_1 = F_2$.

3. Consider the linear model $\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{e}$, where $\boldsymbol{X}$ is $n\times p$ of rank p and $\underline{e} \sim N(0, \sigma^2\mathbf{I})$. Partition $\boldsymbol{X}$ into $\boldsymbol{X} = [\boldsymbol{X}_1|\boldsymbol{X}_2|\boldsymbol{X}_3]$, where $\boldsymbol{X}_i$ is $n_i \times p$ of rank p_i. Equivalently, let $\underline{\beta}^\top = [\underline{\beta}_1^\top|\underline{\beta}_2^\top|\underline{\beta}_3^\top]$. Thus $\underline{Y} = \boldsymbol{X}_1\underline{\beta}_1 + \boldsymbol{X}_2\underline{\beta}_2 + \boldsymbol{X}_3\underline{\beta}_3 + \underline{e}$

   (a) Show that successively fitting the model $\underline{Y} = \underline{e}$, $\underline{Y} = \boldsymbol{X}_1\underline{\beta}_1 + \underline{e}$, $\underline{Y} = \boldsymbol{X}_1\underline{\beta}_1 + \boldsymbol{X}_2\underline{\beta}_2 + \underline{e}$, and $\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{e}$ yields SS's for $\underline{\beta}_1, \underline{\beta}_2$ and $\underline{\beta}_3$ which are orthogonal.

   (b) Prove that the SSE for the model containing $\underline{\beta}$ must be at leat as small as the SSE for the model with only $\underline{\beta}_1$.

   (c) Now, let $\underline{e} \sim N(0, \sigma^2\boldsymbol{V})$, $\widehat{\underline{\beta}}_{ols} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\underline{Y})$ and $\widehat{\underline{\beta}}_{wls} = (\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{V}^{-1}\underline{Y})$. Show that $V(\widehat{\underline{\beta}}_{ols}) - V(\widehat{\underline{\beta}}_{wls})$ is nonnegative definite.

$$\begin{aligned}&(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \geq (\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \\ \Longleftrightarrow\quad &(\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{X}) \leq (\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X}) \\ \Longleftrightarrow\quad &\boldsymbol{X}^\top\{\boldsymbol{V}^{-1} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\}\boldsymbol{X} \geq 0 \\ \Longleftrightarrow\quad &\boldsymbol{V}^{-1} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \geq 0 \\ \Longleftrightarrow\quad &\boldsymbol{I} - \boldsymbol{V}^{1/2}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{V}^{1/2} \geq 0\end{aligned}$$

   Note that $\boldsymbol{I} - \boldsymbol{V}^{1/2}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{V}^{1/2}$ is a symmetric and idempotent matrix. The conclusion follows.

4. Consider two models for Y where for both models e_i are iid $N(0,\sigma^2)$ and $\boldsymbol{X} = [\boldsymbol{X}_1|\boldsymbol{X}_2]$ in $n\times(p_1+p_2)$ of rank (p_1+p_2).

$$M1 : \underline{Y} = \boldsymbol{X}_1\underline{\beta}_1 + \underline{e}$$
$$M2 : \underline{Y} = \boldsymbol{X}_1\underline{\beta}_1 + \boldsymbol{X}_2\underline{\beta}_2 + \underline{e}$$

   (a) Under what conditions does the estimator of $\underline{\beta}_1$ using $M1$ equal the estimator of $\underline{\beta}_1$ using $M2$?

   (b) Prove that the SSE under $M2$ is less than or equal to SSE under $M1$.
   Note that

sth wrong— $\boldsymbol{X}_1^\top\left\{\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{X}_1[\boldsymbol{X}_1^\top(\boldsymbol{I}-\boldsymbol{H}_{x_2})\boldsymbol{X}_1]^{-1}\right\}\boldsymbol{X}_1^\top + \boldsymbol{X}_2\left\{[\boldsymbol{X}_2^\top(\boldsymbol{I}-\boldsymbol{H}_{x_1})\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2^\top\right\}$ **is nonnegative definite.** –HL

$$SSE_2 - SSE_1 = \mathbf{Y}^\top(\mathbf{H}_x - \mathbf{H}_{x_1})\mathbf{Y} = \mathbf{y}\widetilde{\mathbf{X}}_2(\widetilde{\mathbf{X}}_2^\top\widetilde{\mathbf{X}}_2)^{-1}\widetilde{\mathbf{X}}_2\mathbf{Y}$$

   where $\widetilde{\mathbf{X}}_2 = (\mathbf{I}-\mathbf{H}_{x_1})X_2$ (Th7.1 of Ronald, p247).

7 Essential Probablity

Df1.8

(i)A π-class $\mathcal{P}$ is a class of subsets of Ω such that $A, B \in \mathcal{P}$ implies $AB \in \mathcal{P}$

(ii) A semi-field $\mathcal{S}$ is a class of subsets of Ω such that $\mathcal{S}$ is closed under finite intersections, and $A \in \mathcal{S}$ implies $A^c = \cap_{i=1}^m B_i$ where $B_i \in \mathcal{S}$ and disjoint and $m < \infty$.

iii) $\lambda-$class $\mathcal{L} \Longleftrightarrow \Omega \in \mathcal{L}$; $A, B \in \mathcal{L}$ and $A \subset B$ implies $BA^c \in \mathcal{L}$; $A_n \in \mathcal{L}$ and $A_n \uparrow A$ implies $A \in \mathcal{L}$.

Th1.11 (**Dynhin's Class Theorem**) Suppose $\mathcal{P}$ is a $\pi-$class for Ω. Then $\sigma(\mathcal{P}) = \lambda(\mathcal{P})$.

Th1.12 Let $(\Omega, \mathcal{F}_i, \mu_i)$ be measure spaces (i=1,2). Suppose $\mathcal{P}$ is a $\pi-$class such that $\mathcal{P} \subset \mathcal{F}_i$, μ_1 and μ_2 agree on $\mathcal{P}$ and there exist $A_n \uparrow \Omega$ with $A_n \in \mathcal{P}$ and $\mu_i(A_n) < \infty$. Then μ_1 and μ_2 agree on $\sigma(\mathcal{P})$.

Th1.14 (**Carathépdpry's Extension Theorem**) Suppose $\mathcal{F}$ is a field of subsets of Ω and $\mu : \mathcal{P} \to R^+$. If $\mu(\emptyset) = 0$; $\mu(A) \geq 0$ for all $A \in \mathcal{F}$; if $A_i \in \mathcal{F}$ disjoint and $A = \cup A_i$ is in $\mathcal{F}$ then $\mu(A) = \sum\mu(A_i)$. Then there exists a unique extension of μ to $\sigma(\mathcal{F})$.

Th1.16 f is measurable iff $f^{-1}((-\infty, x]) \in \mathcal{F}$ for every $x \in R$ and $f^{-1}(\{-\infty\}) \in \mathcal{F}$, $f^{-1}(\{\infty\}) \in \mathcal{F}$.

Th1.17 f is measurable iff it is the pointwise limit of simple functions.

Th1.18 If f_n are measurable, then $\lim f_n$ is measurable; $f_1 + f_2$ is also measurable; Continuous and monotone functions are **Borel** measurable.

Df1.21 Let $f : \Omega \to \bar{R}$ be $\mathcal{F}$ measurable and μ be σ-finite (i) $f = \sum_{i=1}^m a_i 1_{A_i}$ is simple. Then $\int f d\mu = \sum a_i\mu(A_i)$.
(ii) $f \geq 0$ and $f_n \uparrow f$, where $f_n \geq 0$ is simple. $\int f d\mu = \lim\int f_n d\mu$.
(iii) f is measurable. $\int f d\mu = \int f_+ d\mu - \int f_- d\mu$.

Th1.22 (vii) If $f \geq 0$ a.e. and $\int f d\mu < \infty$ then $f < \infty$ a.e.
(viii) If $f \geq 0$ a.e. and $\mu(\{\omega : f(\omega) > 0\}) > 0$ then $\int f d\mu > 0$.

Th1.24 If g is Riemann integrable on $[a,b]$ then it is Lebesgue integrable on $[a,b]$ (it is also bounded and continuous a.e., i.e. let $A = \{x : x_n \to x \text{ but } g(x_n) \not\to g(x)\}$ then $\mu(A) = 0$. ($[a,b]$ *has to be bounded, otherwise not true. For example,* $\int_0^\infty \frac{\sin x}{x}dx = \pi$ *but* $\int_0^\infty \frac{\sin x}{x}d\mu$ *doesn't exist*)

Th1.26 (**Monotone Convergence**) Suppose $f_n \geq 0$ measurable and $f_n \uparrow f$ a.s., then $\int f_n d\mu \uparrow \int f d\mu$.

Th1.27 (**Fatou's Theorem**) $f_n \geq 0$ measurable, then $\liminf_{n\to\infty}\int f_n d\mu \geq \int \liminf_{n\to\infty} f_n d\mu$.
$f_n \leq f$ integrable, then $\limsup_{n\to\infty}\int f_n d\mu \leq \int \limsup_{n\to\infty} f_n d\mu$.

Th1.28 (**Dominated Convergence Theorem**) $|f_n| \leq g$ and g is integrable. $f_n \to f$ means $\lim_{n\to\infty}\int f_n d\mu = \int f d\mu$.

Th1.29(**Extended DCT**) $|f_n| \leq g_n \to g$, $f_n \to f$. g_n and g integrable and $\lim_{n\to\infty}\int g_n d\mu = \int g d\mu$. Then $\lim_{n\to\infty}\int f_n d\mu = \int f d\mu$.

Th1.33 (**Fubini's Theorem**) $\mu_1 \times \mu_2$ is a product measure on $(\Omega_1\times\Omega_2, \mathcal{F}_1\times\mathcal{F}_2)$. μ_i is σ-finite. $f(w_1,w_2)$ is $\mathcal{F}_1\times\mathcal{F}_2$ measurable and is either non-negative or $\mu_1\times\mu_2$ integrable. Then

$$\begin{aligned}\int_{\Omega_1\times\Omega_2} f(w_1,w_2)d(\mu_1\times\mu_2) &= \int_{\Omega_1}\left\{\int_{\Omega_2} f(w_1,w_2)d\mu_2\right\}\mu_1 \\ &= \int_{\Omega_2}\left\{\int_{\Omega_1} f(w_1,w_2)d\mu_1\right\}\mu_2\end{aligned}$$

Th1.35 Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. If $f, g \geq 0$ and μ is a $\sigma-$finite, then $\int_A f d\mu = \int_A g d\mu$ for all $A \in \mathcal{F} \Longleftrightarrow f = g$ a.e. (μ).
If f, g are $\mu-$ integrable and $\mathcal{P}$ is $\pi-$class generating $\mathcal{F}$, then $\int_A f d\mu = \int_A g d\mu$ for all $A \in \mathcal{P} \Longleftrightarrow f = g$ a.e. (μ).

7.1 Random Variables

Th2.7 $X \sim F$ and $r > 0$. Then $E(|X|^r) = r\int_0^\infty x^{r-1}\{1-F(x)+F(-x)\}dx$.

Th2.10 (**Minkowski Inequality**)
$(\int|f+g|^r d\mu)^{1/r} \leq (\int|f|^r d\mu)^{1/r} + (\int|g|^r d\mu)^{1/r}$

Co2.19 Let $X_1, X_2, \cdots$, be independent random variables (or vectors). If $g_1, g_2, \cdots$ are measurable, then $g_1(X_1), g_2(X_2), \cdots$ are also independent.

$A_1, A_2, \cdots$ independent $\Longleftrightarrow B_1, B_2, \cdots$ independent if each $B_n \in \{\emptyset, A_n, A_n^c, \Omega\}$

Th2.20 Let $F_1, \cdots$ be probability d.f.'s. There exists $(\Omega, \mathcal{F}, P)$ and a sequence of **independent** r.v.'s $X_1, \cdots$ such that $X_i \sim F_i$ for all $i \geq 1$.

7.2 Convergence of Random Variables

Ex3.8 (**A useful lower bound**) Let $Y \geq 0$ with $E(Y^2) < \infty$ and let $a < E(Y)$. Then we have $P(Y > a) \geq (EY - a)^2 / EY^2$. This is often used with $a = 0$.

Proof. Using Cauchy-Schwarz inequality to $Y 1_{(Y>a)}$, we obtain that

$$P(Y > a) \geq \frac{(EY 1_{(Y>a)})^2}{EY^2} = \frac{(EY - EY 1_{(Y \leq a)})^2}{EY^2} \geq \frac{(EY - a)^2}{EY^2}$$

Th3.8

- Suppose X_n, X are a.s. finite and g is *continuous*. Then $X_n \to X$ a.s. (pr) implies $g(X_n) \to g(X)$ a.s. (pr)
- Suppose X_n, Y_n, X, Y are a.s. finite and $X_n \to X$ a.s. (pr) and $Y_n \to Y$ a.s. (pr). Then $X_n + Y_n \to X + Y$ a.s. (pr) and $\max(X_n, Y_n) \to \max(X, Y)$ a.s. (pr)
- If X_n, X, M_n are a.s. finite and $M_n \to \infty$ and $X_n \to X$ a.s. Then $X_{M_n} \to X$ a.s. (*not true for probability convergence*).

Th3.9 (**Borel-Cantelli Theorem**).

- $\sum_{n=1}^{n} P(A_n) < \infty$, then $P(\overline{\lim} A_n) = 0$.
- If A_n are independent, then $\sum_{n=1}^{n} P(A_n) = \infty$ if and only if $P\{\overline{\lim} A_n\} = 1$ or $P\{\underline{\lim} A_n\} = 0$.

Th3.10$X_n \longrightarrow^{pr} X \Longleftrightarrow \forall$ subsequence $n_k \to \infty$ there exist a further subsequence $n_{k_j} \to \infty$ such that $X_{n_{k_j}} \longrightarrow X$ a.s. (*Useful result*)

Th3.11 $X_n \longrightarrow^{pr} X$ and $|X_n| \leq Y$ and $E(Y) < \infty$. Then $E(X_n) \to E(X)$.

Th3.17 (**Glivenko-Cantelli Theorem**) X_n iid $\sim F$ with empirical distribution F_n. Then $P\{\sup_x |F_n(x) - F(x)| \to 0\} = 1$.

Df3.18 (**Tail events**) Let $X_1, \cdots$ be r.v.'s on $(\Omega, \mathcal{F}, P)$. An event A is a tail event for $\{X_n\}$ if $A \in \sigma(X_n, X_{n+1}, \cdots)$ for every n. $\{\lim X_n \text{exits}\}$, $\{\lim X_n = X\}$, $\{\limsup X_n \leq x\}$, $\{\sum^\infty |X_n| < \infty\}$ are all tail events, but $\{\sum^\infty X_n \leq x\}$ not.

Df3.20 $\{X_n\}$ is uniformly integrable if $\sup_n \int_{|X_n|>a} |X_n| dP \to 0$ as $a \to \infty$.

Th3.21 If either of following holds, $\{X_n\}$ is uniformly integrable

(i) $|X_n| \leq Y$ a.s. for each n and $E|Y| < \infty$;

(ii) $|X_n| \leq Y_n$ and $\{Y_n\}$ is uniformly integrable;

(iii) $\sup_n E|X_n|^{1+\varepsilon} < \infty$ for some $\varepsilon > 0$;

(iv) $\sup_n E(|g(X_n)|) < \infty$ and $|g(x)|/|x| \to \infty$ as $|x| \to \infty$

Th3.22 $\{X_n\}$ is uniformly integrable $\Longleftrightarrow$ (i)(**uniformly bounded**) $\sup_n E(|X_n|) < \infty$, and (ii)(**uniformly continuous**) for each $\varepsilon > 0$, there is $\delta > 0$ such that $P(A) < \delta \longrightarrow \sup_n \int_A |X_n| dP < \varepsilon$.

Th3.23 Suppose $0 < p < \infty$ and $X_n \to^{pr} X$. Then the following are equivalent.

(i) $\{X_n\}$ is uniformly integrable

(ii) $X_n \to X$ (L^p) and either $X_n \in L^p$ all n or $X \in L^p$.

(iii) $E(|X_n|^p) \to E(|X|^p) < \infty$

7.3 Convergence of Distributions $\to^{\mathcal{L}}$

Th4.3 $X_n \to^{pr} X$, then $X_n \to^{\mathcal{L}} X$. If $X_n \in (\Omega, \mathcal{F}, P)$, then $X_n \to^{pr} a$ (constant) $\Longleftrightarrow X_n \to^{\mathcal{L}} a$

Th4.4 Let P_n, P be probability measures with densities wrt a common measure μ. $f_n = dP_n/(d\mu)$ and $f = dP/(d\mu)$. If $f_n \to f$ a.e.(μ), then $\sup_{A \in \mathcal{F}} |P_n(A) - P(A)| \leq \int |f_n - f| d\mu \to 0$

Th4.5 (**Slutsky**) $X_n \to^{\mathcal{L}} X$ and $Y_n \to^{\mathcal{L}} C_1$, $Z_n \to^{\mathcal{L}} C_2$. Then $X_n + Y_n \to^{\mathcal{L}} X + C_1$ and $X_n Z_n \to^{\mathcal{L}} C_2 X$. (**Even** $Y_n \to^{pr} Y$ **may** $X_n + Y_n \not\to^{\mathcal{L}} X + Y$)

Th4.9 (**Skorohod's Theorem**) Suppose F_n, F are probability distribution functions on R and $F_n \to^{\mathcal{L}} F$. Then there exists a probability space $(\Omega, \mathcal{F}, P)$ and r.v.'s Y_n, Y such that $Y_n \sim F_n$ and $Y \sim F$ and $Y_n(\omega) \to Y(\omega)$ for **every** $\omega \in \Omega$. (*because of this theorem, some topic about expectation involved into* $\to^{\mathcal{L}}$ *can be transferred into that into* $\to^{a.s.}$)

Th4.10 $X_n \to^{\mathcal{L}} X$

(i) If $X_n \geq 0$ then $\liminf_n E(X_n) \geq E(X)$

(ii) If $P(|X_n| > x) \leq P(|Y| > x)$ and $E(|Y|) < \infty$, then $\lim_n E(X_n) = E(X)$.

(iii) $\{X_n\}$ is uniformly integrable iff $E(X_n) \to E(X)$ as $n \to \infty$.

Th4.11 (**Continuous Mapping Theorem**) Let $X_n \sim F_n$ and $X \sim F$. If $X_n \to^{\mathcal{L}} X$ and $h: R \to R$ measurable and discontinuous only on D with $P(x \in D) = 0$. Then $h(X_n) \to^{\mathcal{L}} h(X)$.

Th4.20(**Taylor expansion of the characteristic function**). If $E|X|^n < \infty$, then $E(X^k) = (-i)^k \phi^{(k)}(0)$ for $k = 1, \cdots, n$ and

$$\phi(t) = \sum_{k=0}^{n-1} \frac{\phi^{(k)}(0)}{k!} t^k + O(t^n) = \sum_{k=0}^{n} \frac{\phi^{(k)}(0)}{k!} t^k + o(t^n) (\text{ as } t \to \infty)$$

Le4.21 $(X, Y) \sim f(x, y)$. Then

$$X + Y \sim \int_{R^1} f(x, z - x) dx, \quad X - Y \sim \int_{R^1} f(x + z, x) dx$$

$$X/Y \sim \int_{R^1} f(xz, x) |x| dx.$$

Th4.22 (**Inverse formula**) For any $x_1 < x_2$,

$$\frac{1}{2}\{F(x_2 + 0) + F(x_2)\} - \frac{1}{2}\{F(x_1 + 0) + F(x_1)\}$$

$$= \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-itx_1} - e^{-itx_2}}{it} \phi(t) dt$$

Th4.26 (**Continuity Theorem**) Let $\{F_n\}$ be a sequence of d.f.'s with c.f.'s ϕ_n. Then $F_n \to^{\mathcal{L}} F$ for some d.f. F$\Longleftrightarrow \phi_n(t) \to \phi(t)$ for all t and ϕ **is continuous at** $t = 0$. In this case, ϕ is the c.f. of F and the convergence $\phi_n \to \phi$ is uniformly in every finite interval.

Th4.28 (**Lindeberg-Lévy central limit theorem**) $\{X_i\}$ are i.i.d. with mean zero and finite variance. Then

$$\lim_{n \to \infty} \left\{ \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^{n} X_i < x \right\} = N(0, 1)$$

De4.30 (**Lindeberg Condition**) A triangular array $\{X_{nk}\}$ with mean zero satisfies the Lindeberg condition if

$$\lim_{n \to \infty} \frac{1}{B_n^2} \sum_{k=1}^{r_n} \int_{|X_{nk}| > \tau B_n} X_{nk}^2 dP = 0 \text{ for any } \tau > 0. B_n^2 = \sum_{i=1}^{r_n} E X_{ni}^2.$$

Th4.31 (**Central limit theorem**) $\{X_i\}$ are independent with finite variance σ_i^2. Then

$$\lim_{n \to \infty} \left\{ \frac{1}{B_n} \sum_{i=1}^{n} (X_i - EX_i) < x \right\} = N(0, 1) \text{ and } \lim_{n \to \infty} \max_{k \leq n} \frac{\sigma_k}{B_n} = 0$$

if and only if *Lindeberg condition* holds. The second one is called **Feller condition** or uniformly asymptotically negligible.

Co4.32 (**Lyapounov theorem**) Suppose $\{X_i\}$ are independent with satisfying, for some $\delta > 0$,

$$\lim_{n \to \infty} \sum_{i=1}^{n} \frac{E|X_i|^{2+\delta}}{\{\sum_{i=1}^{n} \sigma_i^2\}^{1+\delta/2}} = 0$$

Then $\{X_n\}$ satisfies a central limit theorem.

7.4 Absolute continuity/Conditional Expectation

De5.4 Let X be a r.v. on $(\Omega, \mathcal{F}, P)$ such that its mean exists and let $\mathcal{Q}$ be a sub $\sigma-$field of $\mathcal{F}$. The conditional expectation of X given $\mathcal{Q}$ is any r.v. (denote $E(X|\mathcal{Q})$) satisfying

(i) $E(X|\mathcal{Q})$ is measurable wrt $\mathcal{Q}$, and

(ii) $E\{1_B E(X|\mathcal{Q})\} = E(1_B X)$ for any $B \in \mathcal{Q}$.

$$\boxed{E\{g(X)|Y\} = \frac{\int g(x) f(x, Y) dx}{\int f(x, Y) dx}}$$

Th5.9 $\mathcal{P}_1 \subset \mathcal{P}_2$. Then $E\{E(X|\mathcal{P}_1)|\mathcal{P}_2\} = E\{E(X|\mathcal{P}_2)|\mathcal{P}_1\} = E(X|\mathcal{P}_1)$ a.s. In particular, $E\{E(X|Y)|Y, Z\} = E\{E(X|Y, Z)|Y\} = E(X|Y)$ a.s.

An interesting counterexample–Sometimes, $E\{E(Y|X)Z\} \neq E(YZ)$. For example, $Y = g(X) + \varepsilon$. The error term ε is independent of X with zero mean and finite variance σ^2. Let $Z = \varepsilon$. Then $E\{E(Y|X)Z\} = E\{g(X)Z\} = 0$. But $E(YZ) = E(\varepsilon^2) = \sigma^2$

Th5.10 $\mathcal{P}_1$ and $\mathcal{P}_2$ are all $\sigma-$fields and $E\{E(X|\mathcal{P}_1)|\mathcal{P}_2\} = E(X|\mathcal{P}_2 \cap \mathcal{P}_1)$ a.s.

7.5 Related Materials

1. (**Feller-Chung Theorem**) For each integral number j, $A_jA_{j-1}^C\cdots A_0^C$ and B_j are independent, where $A_0=\emptyset$. Then $P(\cup_jA_jB_j)\ge\alpha P(\cup_jA_j)$ for $\alpha=\inf_jP(B_j)$.
   Proof. Note that
   $$\begin{aligned}\cup_jA_jB_j &= A_1B_1+A_2B_2(A_1B_1)^C+\cdots\supset A_1B_1+A_2B_2A_1^C+\cdots\\ P(\cup_jA_jB_j) &\ge P(A_1B_1)+P(A_2A_1^C)P(B_2)+\cdots\ge\alpha\{P(A_1)+P(A_2A_1^C)+\cdots\}\\ &= \alpha P(\cup A_j).\end{aligned}$$
2. $X_1,\cdots,X_n$ are independent and **symmetric** random variables. Write $S_k=\sum_{i=1}^kX_i$ for $k=1,\cdots,n$. Then
   $$P(\max_{1\le k\le n}S_k>a)\le 2P(S_n<a).$$
   Proof. Set $A_k=\{S_1\le a,\cdots,S_{k-1}\le a,S_k>a\}$ and $B_k=\{S_n-S_k\ge 0\}$. Using the results of the previous item, we can prove this conclusion. *This inequality is often used in Wiener processes.*
3. X and Y are independent with means zero. Then $E|X+Y|^r\ge\max(E|X|^r,E|Y|^r)$ for any $r\ge 1$.
   Proof. For any x, $|x|^r=|E(x+Y)|^r\le E|x+Y|^r$. It follows that
   $$\begin{aligned}E|X+Y|^r &= \int|x+y|^rdF_X(x)dF_Y(y)\\ &= \int dF_X(x)\left\{\int|x+y|^rdF_Y(y)\right\}\\ &= \int E|x+Y|^rdF_X(x)\ge\int|x|^rdF_X(x)=E|X|^r.\end{aligned}$$
4. (**Generalized Kolmogorov inequality**) $X_1,\cdots,X_n$ are independent random variables with mean zeros. Write $S_k=\sum_{i=1}^kX_i$ for $k=1,\cdots,n$ and $A=\{\sup_{k\le n}|S_k|\ge C\}$ for some positive constant C. Then
   $$C^rP(A)\le E(|S_n|^rI_A)\le E|S_n|^r\quad\text{for } r\ge 1.$$
   Proof. Set $A_0=\emptyset$ and $A_k=\{\sup_{j<k}|S_j|<C,|S_k|\ge C\}$. Then $A_1,\cdots,A_n$ are disconnect each other and $A=\sum A_k$. It follows from the result of the former item that
   $$E|S_n|^rI_A=\sum_{k=1}^nE|S_n|^rI_{A_k}\ge\sum_{k=1}^nE|S_k|^rI_{A_k}\ge\sum_{k=1}^nC^rEI_{A_k}=C^rP(A),$$
   and we complete the proof of the first assertion. The second one is trivial.
5. If random variable X is integral, then $|median(X)-E(X)|\le\{2var(X)\}^{1/2}$.
6. $E|X|^p\le\infty$ $(p\ge 1)$ if and only if $\sum_{n=1}^\infty\int_{|x|\ge n}|x|^{p-1}dF(x)<\infty$.
7. (**Borel law of large number**) $\mu_n\sim$Bernoulli(n,p), then
   $$P\left\{\lim_{n\to\infty}\frac{\mu_n}{n}=p\right\}=1\quad\Longleftarrow\quad P\left\{\left|\frac{\mu_n}{n}-p\right|>\varepsilon\right\}\le\frac{E|\mu_n-np|^4}{n^4\varepsilon^4}$$
8. (**Hajek-Renyi inequality**) X_i are independent each other with finite variances.
   $$P\left\{\max_{m\le j\le n}\left|C_j\sum_{i=m}^j(X_i-EX_i)\right|>\varepsilon\right\}\le\frac{1}{\varepsilon^2}\left(C_m^2\sum_{i=1}^m\sigma_i^2+\sum_{i=1+m}^nC_i^2\sigma_i^2\right)$$
9. (**Kolmogorov strong law of large number**) X_i are independent each other with finite variances. If $\sum_{n=1}^\infty D(X_n)/b_n^2<\infty$ for $b_n\uparrow\infty$, then $\sum_{i=1}^n(X_i-EX_i)/b_n\to 0$ a.s.
10. (**Kolmogorov strong law of large number**) X_i are i.i.d. $\sum_{i=1}^nX_i/n$ a.s. converges to a if and only if $EX_i<\infty$ and $a=E(X_1)$.
11. Chebyschev inequality, Markov inequality.
12. (**Khintchine law of large number**) X_i i.i.d. with finite mean. Then $\{X_n\}$ is satisfied with law of large numbers.
13. $X_i\sim F_i(x)$. If $\lim_{A\to\infty}\sup_{1\le n<\infty}\int_{|x|>A}|x|dF_n(x)=0$, Then $\{X_n\}$ is satisfied with law of large numbers. (Use **Kolmogorov three series theorem** to prove)
14. X_i are independent. If there exists $\alpha>1$ and $\beta>0$ such that $E|X_n|^\alpha\le\beta$, Then $\{X_n\}$ is satisfied with law of large numbers.
15. (**Markov law of large number**) X_i are independent with mean zero. There exists a $0<\delta\le 1$ such that
    $$\frac{1}{n^{1+\delta}}\sum_{i=1}^nE|X_i|^{1+\delta}\to 0$$
    Then $\{X_n\}$ is satisfied with law of large numbers.
16. (**Elementary Inequality**) $f(x)$ is a **non-decreasing continuous** function. Define
    $$a.e\sup f(\xi)=\inf\{c:P(f|\xi|>c)=0\}.$$
    Then we have the following conclusion:
    $$\frac{Ef(|\xi|)-f(\varepsilon)}{a.e\sup f(\xi)}\le P(|\xi|>\varepsilon)\le\frac{Ef(|\xi|)}{f(\varepsilon)}$$
    In this case, assume $f(0)=0$, then $\xi_n\to 0$ in probability if and only if $Ef(|\xi_n|)\to 0$.
17. (**Gnedenko law of large number**) Taking $f(x)=x^2/(1+x^2)$ in **elementary inequality**, we know that $\{X_n\}$ is satisfied with law of large numbersif and only if
    $$E\left[\frac{\{\sum_{i=1}^n(X_i-EX_i)\}^2}{n^2+\{\sum_{i=1}^n(X_i-EX_i)\}^2}\right]\to 0\Longleftrightarrow\sum_{i=1}^nE\left\{\frac{(X_i-EX_i)^2}{1+(X_i-EX_i)^2}\right\}\to 0\ \text{ if independent}$$
18. $\{X_i\}$ are independent. There exist constants k_n such that $\max_{1\le j\le n}|X_j|\le k_n$ and $k_n/B_n\to 0$. Then $\{X_i\}$ obey central limit theorem.
19. $\{X_i\}$ are independent and obey central limit theorem. $\{X_n\}$ are satisfied with law of large numbersif and only if $B_n^2=o(n^2)$.
20. $\{X_i\}$ are independent. $X_1\sim U[-1,1]$ and $X_k\sim N(0,4^{k-1})$ for $k=2,\cdots$. Then $\{X_k\}$ satisfy central limit theorem (*using c.f. to prove*), but not Lindeberg condition and Feller condition because $b_n/B_n^2\to 1/2$.
21. (**The 1st Helly Theorem**) Suppose $f(x)$ is a continuous function on $[a,b]$, $F_n(x)$ uniformly bound nondecreasing and converge to $F(x)$ on $[a,b]$. a,b are the continuous points of $F(x)$, then
    $$\lim_{n\to\infty}\int_a^bf(x)dF_n(x)=\int_a^bf(x)dF(x).$$
22. (**The 2nd Helly Theorem**) Suppose $f(x)$ is a continuous bounded function on R^1, $F_n(x)$ uniformly bound nondecreasing and converge to $F(x)$ on R^1. In addition $F_n(-\infty)\to F(-\infty)$ and $F_n(\infty)\to F(\infty)$. Then
    $$\lim_{n\to\infty}\int_{-\infty}^\infty f(x)dF_n(x)=\int_{-\infty}^\infty f(x)dF(x).$$

- 147 Suppose X_n are independent, then $X_n\to^{a.s.}0\Longleftrightarrow\forall\varepsilon$, $\sum_{n=1}^\infty P\{|X_n|\ge\varepsilon\}<\infty$.
  The proof can be completed using *Borel-Cantelli Theorem* by setting $A_n=\{|X_n|\ge\varepsilon\}$.
- 179 Suppose $\{X_n,n\ge 1\}$ is iid. Then there exists a sequence of constants C_n such that
  $$\frac{1}{n}\left(\sum_{j=1}^nX_j-C_n\right)\to^{pr}0\Longleftrightarrow\lim_{n\to\infty}nP(|X_1|\ge n)=0$$
  and $C_n=nP(|X_1|\ge n)$. If $E|X_1|<\infty$, C_n can be taken $nE(X_1)$.
- 177 Let $p>0$ and F be the cdf of r.v. X. (i) If $E|X|^p<\infty$, then for any $\alpha>-1,\beta>0$ and $\gamma\ge 0$ satisfying $\frac{\alpha+1}{\beta}+\gamma=p$, we have
  $$\sum_{n=1}^\infty n^\alpha\int_{|x|>n^\beta}|x|^\gamma dF(x)<\infty.\qquad(2)$$
  Conversely, if there is a set of (α,β,γ) satisfying (2), then $E|X|^p<\infty$. (ii) If "$\alpha>-1$" was changed into "$\alpha<-1$", the integral in (2) should also changed as $\int_{|x|\le n^\beta}|x|^\gamma dF(x)$. (*This is a very useful result. For example, prove* **183**. *If $p<1$, taking $\alpha=0,\beta=1/p$ and $\gamma=0$ we have $E|X|^p<\infty\Longleftrightarrow\sum_{n=1}^\infty P(|X|\ge n^{1/p})<\infty$; If $p>1$, taking $\alpha=0,\beta=1$ and $\gamma=p-1$ we have $E|X|^p<\infty\Longleftrightarrow\sum_{n=1}^\infty E\{|X|^{p-1}I_{|X|\ge n}\}<\infty$*)
- 183 (**Marcinkiewicz-Zygmund Theorem**) Suppose $\{X_n,n\ge 1\}$ is iid and $p\in(0,2)$. Then there exists a sequence of constants C_n such that
  $$n^{-1/p}\left(\sum_{j=1}^nX_j-C_n\right)\to^{a.s.}0\Longleftrightarrow E|X_1|^{1/p}<\infty,$$
  and $C_n=0$ if $0<p<1$ and $nE(X_1)$ otherwise.
- 211 Suppose $\{X_n,n\ge 1\}$ is iid. Then
  $$\frac{1}{n}\max_{1\le i\le n}X_i\to^{pr}0\Longleftrightarrow nP(|X_1|>n)=o(1)\text{ and }\frac{1}{n}\max_{1\le i\le n}X_i\to^{a.s}0\Longleftrightarrow E|X$$
- (**Doob Inequality**) For independent sequence $\{X_n\}$ with mean zero and $p>1$,
  $$E\left(\max_{1\le k\le n}\left|\sum_{j=1}^kX_j\right|^p\right)\le\left(\frac{p}{p-1}\right)^pE\left(\left|\sum_{j=1}^nX_j\right|^p\right)$$
- **65, 81, 138** in the big notebook
- Characteristic Functions

Dis.	density	c.f.	additivity
$B(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$(pe^{it}+1-p)^n$	$B(n_1,p)*B(n_2,p)=B(n_1+n_2,p)$
$P(\lambda)$	$\frac{\lambda^x}{x!}\exp(-\lambda)$	$\exp\{\lambda(e^{it}-1)\}$	$P(\lambda_1)*P(\lambda_2)=P(\lambda_1+\lambda_2)$
$N(\mu,\sigma^2)$	$(\sqrt{2\pi}\sigma)^{-1}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	$\exp(i\mu t-\frac{\sigma^2t^2}{2})$	$N(\mu_1,\sigma_1^2)*N(\mu_2,\sigma_2^2)=N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$
$\Gamma(\alpha,\lambda)$	$\frac{\alpha^\lambda}{\Gamma(\lambda)}x^{\lambda-1}e^{-\alpha x}$	$\left(1-\frac{it}{\alpha}\right)^{-\lambda}$	$\Gamma(\alpha,\lambda_1)*\Gamma(\alpha,\lambda_2)=\Gamma(\alpha,\lambda_1+\lambda_2)$
$C(\alpha,\mu)$	$\frac{\alpha}{\pi\{\alpha^2+(x-\mu)^2\}}$	$e^{i\mu t-\alpha\|t\|}$	$C(\alpha_1,\mu_1)*C(\alpha_2,\mu_2)=C(\alpha_1+\alpha_2,\mu_1+\mu_2)$
$\chi^2(n)$	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}$	$(1-2it)^{-\frac{n}{2}}$	$\chi^2(n_1)*\chi^2(n_2)=\chi^2(n_1+n_2)$

8 Basics of Statistical Inference

Definition 1 *Suppose X is an observation from an unknown distribution $P \in \mathcal{P}$ where $\mathcal{P}$ is a family of distributions. A statistic $T = T(X)$ is said to be sufficient for $P \in \mathcal{P}$ (or $\theta \in \Theta$ where $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is a parametric family) if the conditional distribution of X given T does not depend on P (or θ in parametric family).*

Theorem 1 *(**N-F factorization Theorem**). Suppose X is an observation from $P \in \mathcal{P} = \{f(\cdot,\theta), \theta \in \Theta\}$. A statistic $T = T(X)$ is sufficient for $\mathcal{P}$ **iff** $f(x,\theta) = g(t(x),\theta)h(x)$ where g and h re two functions and h if free of θ.*

Theorem 2 *Consider the model $\mathcal{P} = \{f(\cdot,\theta); \theta \in \Theta\}$. Let T be a minimal sufficient statistic (T is a function of any sufficient statistic). Then for any x, y in the sample space.*

$$T(x) = T(y) \iff L(\theta,x) \propto L(\theta,y)$$

Remark 1
- *Theorem 2 shows that the likelihood function is equivalent to the following: Sufficiency principle inference for a given model which admits a minimal sufficient statistic T should be identical for any x and y such that $T(x) = T(y)$.*
- *Thereon 2 also provides methods for identifying minimal sufficient statistic. If T is a statistic such that for any x, y*

  $$T(x) = T(y) \iff L(\theta,x) \propto L(\theta,y).$$

  Then T is minimal sufficient.

Definition 2 *The family $\{f(\cdot,\theta); \theta \in \Theta\}$ constitutes an exponential family if*

$$f(x,\theta) = h(x)\exp\left\{\sum_{i=1}^{k} \phi_i(\theta)T_i(x) - \tau(\theta)\right\}$$

where $T_1, \cdots, T_k$ and h are functions of x not depending on θ, and $\phi_1, \cdots, \phi_k$ and τ are functions of θ not depending on x.

Theorem 3 *If the exponential family is in reduced form, then $(T_1(x), \cdots, T_k(x))$ is **minimal sufficient statistic**.*

Theorem 4 *If $k = 1$ and $f(x,\theta) = h(x)\exp\{\phi(\theta)T(x) - \tau(\theta)\}$, then the moments of T can be represented in terms of ϕ and τ*

$$E_\theta\{T(X)\} = \frac{\tau'(\theta)}{\phi'(\theta)} \quad \text{and} \quad Var_\theta\{T(X)\} = \frac{\phi'(\theta)\tau''(\theta) - \tau'(\theta)\phi''(\theta)}{\phi'^3(\theta)}.$$

and so on.

Definition 3 *A statistic $T(x)$ is an ancillary statistic if its distribution does not depend on the parameter.*

Definition 4 *Suppose $\mathbf{x}$ is an observation from an unknown distribution $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$. A statistic $T = T(X)$ is said to be complete if for any measurable function g, $Eg(T) = 0$ for all $P \in \mathcal{P}$ means that $g(t) = 0$ a.e. P*

We say that $T = T(X)$ is boundedly complete if the previous statement holds for and bounded measurable function g.

Example 1 *(a) Suppose $X \sim B(n,p)$. Suppose g is such that $\sum_{x=0}^{n} g(x)\binom{n}{x}p^x(1-p)^x = 0$ for all p. This means $g(x) = 0$ for $x = 0, 1, \cdots, n$, and X is complete.*

(b) $X_1, \cdots, X_n$ are iid $U(0,\theta)$ for $\theta \in (0,\infty)$. Then $X_{(n)}$ is sufficient. Suppose g is such that $0 = Eg(X_{(n)}) \propto \int g(t)t^{n-1}dt$ for all θ. This means that $g(t)t^{n-1} = 0$ a.e. from measure theory. Thus $X_{(n)}$ is complete.

(c) $X_1, \cdots, X_m$ are iid with pdf $N(\mu,\sigma_1^2)$ and $Y_1, \cdots, Y_n$ are iid with pdf $N(\mu,\sigma_2^2)$ and they are independent. $T = (\sum_{i=1}^m X_i^2, \sum_{i=1}^n Y_i^2, \sum_{i=1}^n Y_i, \sum_{i=1}^m X_i)$ is sufficient. But T is not complete since $E\sum X_i - E\sum Y_i = 0$ for all $\mu, \sigma_1^2, \sigma_2^2$.

(d) $X_1, \cdots, X_n$ are iid $U(\theta, \theta+1)$ for $\theta \in R^1$. Then $T = (X_{(1)}, X_{(n)})$ is minimal sufficient but not complete(HW).

(e) Suppose X is observation from

$$P \in \left\{f(x,\phi_1,\cdots,\phi_k) = h(x)\exp\{\sum_{i=1}^{k}\phi_i T_i(x) - \xi(\phi)\}, \phi \in \Theta\right\}$$

*where Θ contains an open set (the family is said to be full-rank). Then $(T_1, \cdots, T_k)$ is **complete**.*

Theorem 5 *Let T be a one-dimensional complete and sufficient statistic. Then it is minimal sufficient.*

Lemma 1 *Let X, Y be r.v.'s where Y has finite variance, then (i) $E(E(Y|X)) = E(Y)$ and (ii) $Var\{E(Y|X)\} \le Var(Y)$.*

Theorem 6 *(**Basu Theorem**) Let V and T be two statistics based on an observation X from $P_\theta \in \mathcal{P}$. If T is boundedly complete and sufficient and the distribution of V doesn't depend on θ. Then V and T are independent for any θ.*

8.1 Point Estimation

Theorem 7 *(**Blackwell-Lehmann-Rao-Scheffé Theorem**)(B-L-R-S) Let X be an observation from a distribution in a family $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$. Assume that $g(\theta)$ is U-estimable and U is an unbiased estimator of $g(\theta)$.*

- *If T is sufficient for $g(\theta)$, $E(U|T)$ is also unbiased for $g(\theta)$ and $Var_\theta[E(U|T)] \le Var_\theta(U)$ for all θ.*
- *If T is complete and sufficient for θ, Then $E(U|T)$ is the unique UMVUE of $g(\theta)$. (Here unique means if there exists another estimator V which is UMVUE, then $V(x) = U(X)$ a.e.P_θ)*
- Therefore if a unbiased estimator is not one function of a complete and sufficient statistic, the estimator must not be UMVUE. For example the sample variance S_n^2 for σ^2 in $N(0,\sigma^2)$.

The following are typical approaches for deriving UMVUE when a complete and sufficient statistic T is available.

> (i) We happen to know that $\phi(T)$ is unbiased for $g(\theta)$, then $\phi(T)$ is UMVUE of $g(\theta)$.
>
> (ii) We first identify an unbiased estimator U of $g(\theta)$, and then calculate $E(U|T)$, which is UMVUE.
>
> (iii) In some case, one can solve $E_\theta\phi(T) = g(\theta)$ for ϕ.

Theorem 8 *Let $\mathcal{U}$ be the set of all unbiased estimators of 0 with finite variance and T an unbiased estimator of $g(\theta)$. A necessary and sufficient condition for T to be UMVUE is that $Cov(U,T) = 0$ for all $U \in \mathcal{U}$.*

Theorem 9 *(**The Cramér-Rao Lower Bound**). Let X be an observation from $P \in \mathcal{P} = \{P_\theta, \theta \in \Theta\}$, where Θ is an open set in R^k. Suppose that $T = T(X)$ is an unbiased estimator of $g(\theta)$, where g is differentiable at all $\theta \in \Theta$. Further, suppose that P_θ has a density function $f(x,\theta)$ w.r.t some measure ν for all $\theta \in \Theta$, and $f(x,\theta)$ is differentiable in θ and satisfies that*

$$\frac{\partial}{\partial\theta}\int h(x)f(x,\theta)d\nu = \int h(x)\frac{\partial}{\partial\theta}f(x,\theta)d\nu \tag{3}$$

for all $\theta \in \Theta$, and $h(x) = 1$ and $h(x) = T(x)$. Then

$$Var_\theta\{T(X)\} \ge \left(\frac{\partial}{\partial\theta}g(\theta)\right) I^{-1}(\theta) \left(\frac{\partial}{\partial\theta}g(\theta)\right)^T$$

where $I(\theta) = E_\theta\left\{\left(\frac{\partial}{\partial\theta}\log f(X,\theta)\right)\left(\frac{\partial}{\partial\theta}\log f(X,\theta)\right)^T\right\}$

The r.v. $\frac{\partial}{\partial\theta}f(x,\theta)$ is called *the efficient score* of θ. $I(\theta)$ is called Fisher Information Matrix.

Series Expansion Method Often we wish to estimate $g(\theta)$ when we have an unbiased estimator T of θ. We are attempted to use $g(T)$ as the estimator of $g(\theta)$, but this is typically biased. We can express $g(T)$ about θ using Taylor series

$$g(T) \approx g(\theta) + g'(\theta)(T-\theta) + \frac{1}{2}g''(\theta)(T-\theta)^2$$

Taking expectation both side, we get

$$Eg(T) \approx g(\theta) + \frac{1}{2}g''(\theta)Var(T).$$

Often $Var(T) = O(1/n)$. This means that the bias has order $1/n$. In some cases, we can estimate $g''(\theta)Var(T)$ and modify $g(T)$ accordingly. So that it will have smaller bias.

Jackknife. $\overline{T}_{n-1,\cdot} = \frac{1}{n}\sum_{j=1}^n T_{n-1,j}$. Define $\boxed{T_n^J = nT_n - (n-1)\overline{T}_{n-1,\cdot}}$.

8.2 Maximum Likelihood Estimation (MLE)

Definition 5 *Suppose $\mathbf{X}$ is a sample from $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ where P_θ is assumed to have a density $f(x,\theta)$. Let $L(\theta,x)$ be the likelihood function. A statistic $\widehat{\theta} \in \Theta$ satisfying*

$$L(\widehat{\theta},\mathbf{X}) = \max_{\theta\in\Theta} L(\theta,\mathbf{X})$$

is called maximum likelihood estimate of θ. $\widehat{\theta}$ viewed as an estimator is called maximum likelihood estimator.

Newton-Raphson Let θ_0 be a fixed point. Write $Dl(\theta,x) = \left[\frac{\partial}{\partial\theta_1}l(\theta,x), \cdots, \frac{\partial}{\partial\theta_k}l(\theta,x)\right]^T$ and $D^2l(\theta,x) = \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}l(\theta,x)\right]_{i,j=1\sim k}$. $\widehat{\theta} \approx \theta_0 - [D^2l(\theta_0,x)]^{-1}Dl(\theta_0,x)$.

EM-algorithm
E-step (Estimation step). Compute $Q(\theta|\theta_k) = E_{\theta_k}[\log L(\theta,X)|Y]$;
M-step (Maximization step). Select θ_{k+1} as the maximization of $Q(\theta|\theta_k)$.
Apply these steps iteratively until "convergence".

Theorem 10 *Let $X_1, \cdots, X_n$ be iid with a common density $f(x,\theta)$ w.r.t a σ-finite measure (focus on pdf and pmf) where θ is real-valued. Assume the following conditions.*

(a) The parameter space θ is an open interval (finite or infinite)

(b) The distribution P_θ of X_i have common support so that $A = \{x, f(x,\theta) > 0\}$ is independent of θ.

(c) For any $x \in A$, the density $f(x,\theta)$ is three times differentiable in θ,

(d)

$$E_\theta\left[\frac{\partial}{\partial\theta}\log f(X,\theta)\right] = 0.$$

and

$$E_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X,\theta)\right] = -I(\theta).$$

(e) There exists a finite neighbor $c(\theta_0-\varepsilon,\theta_0+\varepsilon)$ and a function $M(x)$ such that

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x,\theta)\right| \le M(x)$$

for all $x \in A$ and $\theta \in c(\theta_0-\varepsilon,\theta_0+\varepsilon)$ with $E_{\theta_0}M(X) < \infty$.

Then any consistent sequence $\widehat{\theta}_n$ of roots of the likelihood equation satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \longrightarrow N(0, I^{-1}(\theta)).$$

Definition 6 *Let $X_1, \cdots, X_n$ be iid with a common density $f(x,\theta)$. Let $I(\theta)$ be the Fisher information of X, which is assumed to be well-defined and finite. Let T_n be an estimator of $g(\theta)$ where $g(\theta)$ is differentiable with $g'(\theta) > 0$. We say T_n is* **asymptotically efficient** *if*

$$\sqrt{n}(T_n - g(\theta)) \longrightarrow N(0, v(\theta))$$

where $v(\theta) = [g'(\theta)]^2/(I(\theta))$.

Theorem 11 *Suppose that the conditions of Theorem 10 hold and that $\widetilde{\theta}_n$ is any root-n consistent estimators, i.e. $\sqrt{n}(\widetilde{\theta}_n - \theta_0) = O_p(1)$. Then the estimator sequence*

$$T_n = \widetilde{\theta}_n - \frac{l'(\widetilde{\theta}_n, x)}{l''(\widetilde{\theta}_n, x)}$$

is asymptotically efficient. (In fact $n^{1/4}(\widetilde{\theta}_n - \theta_0) = o_p(1)$ is enough)

8.3 Robustness

Stieltiess Integral *$F(x) = pF_c(x) + qF_d(x)$, where $F_c(x)$ has a derivative $f_c(x)$ and $F_d(x)$ is a step function with discontinuous points $x_1, \cdots, x_n$, the size of jump at the x_i is p_i. Then*

$$\int g(x)dF(x) = p\int g(x)dF_c(x) + q\sum_{i=1}^n g(x_i)p_i.$$

Definition 7 *The influence function (or curse) of T at F is defined for each x by*

$$IF(x,T,F) = \lim_{\varepsilon\downarrow 0}\frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon}$$

when the limit exists.

It is usually true that $\sqrt{n}\{T(F_n) - T(F)\} \longrightarrow N\left\{0, \int IF^2(x,T.F)dF(x)\right\}$.

8.4 Decision Theoretic Estimation

- Risk function:
  $R_T(\theta) = \int L(T(x), g(\theta))p_\theta(x)d\mu(x) = E_\theta[L(T(x), g(\theta))]$. An estimator T is said to be admissible if there is no other estimator T^* such that $R_{T^*}(\theta) \le R_T(\theta)$ for all $\theta \in \Theta$ and $R_{T^*}(\theta) < R_T(\theta)$ for some θ.

- Bayes Risk: $\overline{R}_T = \int R_T(\theta)\pi(\theta)d\theta$. An estimator T_π is said to be **Bayes** estimator with respect to π if $\overline{R}_{T_\pi} \le \overline{R}_T$ for all estimator T.

  Special Cases:

  - $L(T(x), g(\theta)) = \{T(x) - g(\theta)\}^2$.
    $\rho_T(x) = \int\{T(x) - g(\theta)\}^2\pi(\theta|x)d\theta$. Thus $\rho_T(x)$ is minimized by $T(x) = E\{g(\theta)|X\}$.

  - $L(T(x), g(\theta)) = |T(x) - g(\theta)|$.
    $\rho_T(x) = \int|T(x) - g(\theta)|\pi(\theta|x)d\theta$. Thus $\rho_T(x)$ is minimized by $T(x) =$ the **median** of $g(\theta)$ given X, which is just Bayes estimator.

  - $L(T(x), g(\theta)) = w(\theta)\{T(x) - g(\theta)\}^2$.
    $\rho_T(x) = \int w(\theta)\{T(x) - g(\theta)\}^2\pi(\theta|x)d\theta$. Thus $\rho_T(x)$ is minimized by

    $$T(x) = \frac{E\{w(\theta)g(\theta)|X\}}{E\{w(\theta)|X\}}.$$

- An estimator T_π is said to be **Minimax** estimator if $\sup_{\theta\in\Theta}R_T(\theta) \le \sup_{\theta\in\Theta}R_{T^*}$ for any other estimator T^* of θ.

  Suppose T_π is Bayes with respect to π and T_π has constant risk. Then T_π is minimax.

Remark: The MLE or UMVUE may be inadmissible!

Proposition 8.1 *If T_π is unique Bayes with respect to prior π, then T_π is admissible.*

8.5 Hypothesis Test

- parameter space Θ
- Hypothesis $\quad H_0: \theta \in \Theta_0 \subset \Theta \Longleftrightarrow H_1: \theta \in {\Theta_0}^c$
- Reject H_0 if $\delta(x) = 1$ and accept (do not reject) H_0 if $\delta(x) = 0$.
- $\gamma(\theta) = P_\theta(\delta(x) = 1)$ is called **power function** of δ at θ.
- $\alpha(\theta) = \gamma(\theta)$=probability of type I error for $\theta \in \Theta_0$;
- $\beta(\theta) = 1 - \gamma(\theta)$=probability of type II error for $\theta \in {\Theta_0}^c$.

1. (UMT) A test φ of size α is a uniformly most powerful test if $\gamma_\varphi(\theta) \ge \gamma_{\tilde{\varphi}}(\theta)$ for all $\theta \in {\Theta_0}^c$ and size of α.

2. (**N-P Lemma**) Any type of the form $\varphi(x) = \begin{cases} 1 & f_1/f_0 > c \\ \xi(x) & f_1/f_0 = c \\ 0 & o.w \end{cases}$

   for some $c \ge 0$ and $0 \le \xi(x) \le 1$ satisfying $E_{\theta_0}\{\varphi(X)\} = \alpha$ is MP-level $-\alpha$ test.

3. Let $\{f_\theta; \theta \in \Theta\}$ be a family with **MLR** in $T(x)$. (*for all $\theta < \theta'$, $f_{\theta'}(x)/f_\theta(x)$ is a non-decreasing function of $T(x)$*)

   (i) For testing $H_0: \theta \le \theta_0$ vs $H_1: \theta > \theta_0$, there exists a UMP test of level α given by

   $$\varphi(x) = \begin{cases} 1 & T(X) > c \\ \xi & T(X) = c \\ 0 & o.w \end{cases} \tag{4}$$

   where c and ξ are determined by

   $$E_{\theta_0}\varphi(X) = \alpha \tag{5}$$

   (ii) The power function $\gamma(\theta) = E_\theta\varphi(X)$ of the test (4) is strictly increasing for all θ.

   (iii) For all θ', the test determined by (4) and (5) is UMP for testing $H_0: \theta \le \theta'$ vs $H_1: \theta > \theta'$ at level $\alpha' = \gamma(\theta')$.

4. Suppose $X = (X_1, \cdots, X_n)$ is a random variable from the **one-dimensional exponential** family

   $$f_\theta(x) = h(x)\exp\{T(x)\theta - \tau(\theta)\}$$

   then the **UMPU** test for $H_0: \theta = \theta_0 \Longleftrightarrow H_1: \theta \ne \theta_0$ is given by

   $$\varphi(x) = \begin{cases} 1 & T(x) < C_1 \text{ or } T(X) > C_2 \\ \xi_i & T(x) = C_i \text{ for } i = 1,2 \\ 0 & C_1 < T(x) < C_2 \end{cases} \tag{6}$$

   where C_i and ξ_i are determined by the following two equations

   $$E_{\theta_0}\{\varphi(X)\} = \alpha \qquad E_{\theta_0}\{T(X)\varphi(X)\} = \alpha E_{\theta_0}\{T(X)\}.$$

5. Suppose X has pdf $f_{\theta,\eta}(x) = c(\theta,\eta)h(x)\exp\left\{\theta u(x) + \sum_{i=1}^k \eta_i T_i(x)\right\}$ where $(\theta,\eta) \in R^{k+1}$. Define

   $$\varphi(u,t) = \begin{cases} 1 & u > c(t) \\ \xi(t) & u = c(t) \\ 0 & o.w \end{cases}$$

   where $c(t)$ and $\xi(t)$ are determined by

   $$E_{\theta_0}[\varphi\{U(X), T(X)\}|T(X) = t] = \alpha$$

   for all t with $T = (T_1, \cdots, T_k)$. Then φ is a **UMPU** level $1-\alpha$ test for $H_0: \theta \le \theta_0 \Longleftrightarrow H_1: \theta > \theta_0$.

6. Assume the setting as in item 5 and consider testing $H_0: \theta = \theta_0 \Longleftrightarrow H_1: \theta \ne \theta_0$. Then the **UMPU** test of level α is given by

   $$\varphi(u,t) = \begin{cases} 1 & u > c_1(t) \text{ or } u < c_2(t) \\ \xi_i(t) & u = c_i(t) \\ 0 & o.w \end{cases}$$

   where $c_i(t)$ and $\xi_i(t)$ are determined by $E_{\theta_0}\{\varphi\{U,T\}|T = t\} = \alpha$ and $E_{\theta_0}[U\varphi\{U,T\}|T = t] = \alpha E_{\theta_0}\{U|T = t\}$ for all t.

7. Suppose X has pdf $f_{\theta,\eta}(x) = c(\theta,\eta)h(x)\exp\left\{\theta u(x) + \sum_{i=1}^k \eta_i T_i(x)\right\}$ and that $V = V(u,T)$ is independent of T when $\theta = \theta_0$.

(a) Assume further that $V(u,t)$ is increasing in u for each fixed t. Then the **UMPU** test of $H_0 : \theta \le \theta_0 \Longleftrightarrow H_1 : \theta > \theta_0$ is given

$$\varphi(u,t) = \begin{cases} 1 & V > C \\ \xi & V = C \\ 0 & o.w \end{cases}$$

where C and ξ are determined by $E_{\theta_0}\{\varphi(X)\} = \alpha$.

(b) Assume further that $V(u,t) = a(t)u + b(t)$ where $a(t) > 0$ for all t. Then the **UMPU** test of $H_0 : \theta = \theta_0 \Longleftrightarrow H_1 : \theta \neq \theta_0$ is given

$$\varphi(V) = \begin{cases} 1 & V < C_1 \text{ or } V > C_2 \\ \xi_i & V = C_i \\ 0 & o.w \end{cases}$$

where C_i and ξ_i are determined by $E_{\theta_0}\{\varphi(V)\} = \alpha$ and $E_{\theta_0}\{\varphi(V)V\} = \alpha E_{\theta_0}(V)$.

8. (**LRT**) $\lambda(x) = \dfrac{\sup_{\theta\in\Theta} f_\theta(x)}{\sup_{\theta\in\Theta_0} f_\theta(x)}$.
9. **Wald** test statistic
$W = n\{R(\widehat{\theta})\}^T \left[\{\frac{\partial}{\partial\theta}R(\widehat{\theta})\}^T I^{-1}(\widehat{\theta})\{\frac{\partial}{\partial\theta}R(\widehat{\theta})\}\right]^{-1} R(\widehat{\theta})$
$(H_0 : R(\theta) = 0 \Longleftrightarrow H_1 : R(\theta) \neq 0)$
10. **Rao's** score statistic $(H_0 : \theta = \theta_0 \Longleftrightarrow H_1 : \theta \neq \theta_0)$
$S = \frac{1}{n}\left\{\frac{\partial l(\theta_0)}{\partial\theta}\right\}^T I^{-1}(\theta_0)\left\{\frac{\partial l(\theta_0)}{\partial\theta}\right\}$.
11. For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the *acceptance* region of a level α test of $H_0 : \theta = \theta_0$. For each $x \in \mathcal{X}$, define $C(x) = \{\theta_0 : x \in A(\theta_0)\}$. Then random set $C(X)$ is a $1-\alpha$ *confidence set*. Vice versa.
12. A $1-\alpha$ **highest posterior density** (HPD) credible set for θ is a subset $\mathcal{C}$ of θ, of the form $\mathcal{C}_\alpha = \{\theta \in \Theta : \pi(\theta|x) > K(\alpha)\}$ where $K(\alpha)$ is the largest const s.t. $P(\mathcal{C}_\alpha|x) \ge 1-\alpha$.

8.6 Standard Errors

1. $\text{Se}(\widehat{\theta}) = \{E(\widehat{\theta}-\theta)^2\}^{1/2}$
2. $\text{Rmse}(\widehat{\theta}) = \left\{\frac{1}{r}\sum_{j=1}^r(\widehat{\theta}_j-\theta)^2\right\}^{1/2}$, where $\{\widehat{\theta}_j\}$ are the estimators $\widehat{\theta}$ for r replications
3. $\text{Ese}(\widehat{\theta})$: estimate $\text{Se}(\widehat{\theta})$, when it depends another parameters, e.g. $\text{Se}(\overline{X}) = \sigma/n$, then $\text{Ese}(\widehat{\theta}) = S/\sqrt{n}$.
4. $\text{Ase}(\widehat{\theta})$
5. $\text{Ease}(\widehat{\theta})$: estimate $\text{Ase}(\widehat{\theta})$
6. All of above assume that one knows the PDF of what one is sampling from
7. $\text{Jse}(\widehat{\theta})$: Jacknife $\text{Se}(\widehat{\theta})$

$$Jse(\widehat{\theta}) = \left[\frac{n-1}{n}\sum_{i=1}^n\{\widehat{\theta}_{(i)} - Je(\widehat{\theta})\}^2\right]^{1/2} \quad \text{where } Je(\widehat{\theta}) = \overline{\widehat{\theta}}_{(i)}$$

$\overline{\widehat{\theta}}_{(i)} = \widehat{\theta}$ with the $i-$th obs left over.

8. $\text{Bse}(\widehat{\theta})$: Bootstrap $\text{Se}(\widehat{\theta})$

$$Bse(\widehat{\theta}) = \left[\frac{1}{B}\sum_{b=1}^B\{\widehat{\theta}_b - be(\widehat{\theta})\}^2\right]^{1/2} \quad \text{where } be(\widehat{\theta}) = \frac{1}{B}\sum_{b=1}^B\widehat{\theta}_b$$