BI476: Biostatistics - Case Studies

Maoying, Wu
ricket.woo@gmail.com

Dept. of Bioinformatics & Biostatistics
Shanghai Jiao Tong University

Spring, 2018

Syllabus

Calculus

Linear algebra

Optimization

Next Section ...

BI476: Syllabus

- This is a one-semester course for undergraduate students majored in biostatistics or bioinformatics.
- Topics will cover experiment design, intuitive hypothesis testing, (generalized linear models (including generalized estimating equations), survival analysis and multivariate statistics methods.
- Advanced topics will be included, such as penalized regression or hierarchical /mixed-effects linear models or Bayesian statistics, if time permits.
- Estimation, interpretation, and diagnostic approaches will be discussed.
- Software instruction will be provided in class in R.
- Performance will be evaluated based on homeworks (35%), two exams (30%), lab assignments (20%) and projects (15%).
- This is a two-credit course.

BI476: Syllabus

- Textbooks: No textbook is required. We will provide the readings and related materials on the website.
- Prerequisites:Linear algebra, Probability, Biostatistics
- Course Objectives: Upon successful completion of the course, the student will be able to
  - Design the experiment and analyze the experiment data
  - Apply, interpret and diagnose linear regression models
  - Apply, interpret and diagnose logistic, poisson and Cox regresssion models
  - Apply and interpret the multivariate analysis methods.

BI476: Syllabus

Instructor: Maoying Wu (ricket.woo@gmail.com)
Website: `http://cbb.sjtu.edu.cn/~mywu/bi476`
Github: `https://github.com/ricket.sjtu/bi476`
Office: Biopharmaceutics Building, Rm 4A-223
Time: Mondays, 14:00-15:40
Location: East Lower Hall, Rm 403
Office Hours: Monday-Friday, 8:30 17:00

BI476: Schedule

- Lecture 1: Recaps of Mathematical Knowledge for Biostatistics (1-2)
- Lecture 2: Observational Studies and Analysis (3-4)
- Lecture 3: Randomized Clinical Trials and Analysis (5-6)
- Lecture 4: Linear Regression Models and Extensions (7-8)
- Midterm (9)
- Lecture 5: Panel Data Analysis (10-11)
- Lecture 6: Survival Analysis and Competing Risks (12-13)
- Lecture 7: Multivariate Statistcal Analysis (14-15)
- Projects (16)
- Final (17-18)

List of Symbols

- $\mathcal{X}$: A set
- $\mathbb{R}$: Real set
- X: A random variable
- $\mathbf{X} \in \mathbb{R}^{m \times n}$: A m-by-n real matrix
- $\mathbf{x} \in \mathbb{R}^n$: A real vector of length n
- $x \in \mathbb{R}$: A real number
- $\Phi(x) = P(X \leq x)$: Cumulative distribution function
- $f(x) = f(X = x)$: Probability density function

Lecture 1: Mathematics for Biostatistics

- Calculus (微积分)
  - Limits (极限)
  - Derivatives (导数): First-order and second-order
  - Integration (积分)
  - Gradient (梯度)：Jacobian, Hessian
  - Convex function (凸函数) and Jensen's inequality (简森不等式)
  - Taylor's expansion (泰勒展开)
- Linear algebra (线性代数)
  - Vector (向量), matrix (矩阵)
  - Norm (范数)
  - Rank (秩), determinant (行列式), trace (迹)
  - Matrix multiplication (矩阵乘积)
  - Eigendecomposition (正交分解)
  - Singular value decomposition (SVD，奇异值分解)

Lecture 1: Mathematics for Biostatistics

- Probability (概率论)
  - Probability density function (pdf, 概率密度函数)
  - Probability mass function (pmf, 概率质量函数)
  - Cumulative distribution function (cdf, 累积分布函数)
  - Moment generating function (mgf, 矩母函数)
  - Joint probability distribution (联合概率分布)
  - Conditional probability distribution (条件概率分布)
  - Marginal distribution (边缘概率分布)
  - Bayes' Equation/Theorem (贝叶斯公式/定理)
  - Continuous distributions (连续概率分布)
  - Discrete distribution (离散概率分布)
- Numerical Optimization (数值优化方法)
  - Convex set (凸集)
  - Convex function (凸函数), Concave function (凹函数)
  - Gradient descent (梯度下降), gradient ascent (梯度上升)
  - Newton's method (牛顿法)
  - Quasi-Newton's methods (拟牛顿法)
  - Method of multiplers (乘子法)
  - Laglangian method of multiplier (拉格朗日乘子法)

Next Section ...

Limit (极限)

Compute the limit

$$\lim_{x \to 0} \frac{3\sin^2 x}{4x^2}$$

Derivative (导数)

$$f'(x) = \frac{d}{dx}f(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

The derivative is the slope of the tangent line to the graph of $f(x)$, assuming the tangent line exists.

Common Derivatives and Rules

- $\frac{d}{dx}x^n = nx^{n-1}$
- $\frac{d}{dx}\ln = \frac{1}{x}$
- $\frac{d}{dx}a^x = a^x \ln a$
- $\frac{d}{dx}e^x = e^x$

Product rule

$$[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x)$$

Quotient rule

$$\left[\frac{f(x)}{g(x)}\right]' = \frac{g(x)f'(x) - f(x)g'(x)}{g^2(x)}$$

Chain rule

Let $y = f(g(x))$,

$$y' = f'(g(x))g'(x)$$

Integrals (积分)

For a function $f(x)$, its indefinite integral is:

$$\int f(x)dx = F(x) + C, \text{where } F'(x) = f(x)$$

Common integrals and rules

(a) $\int_a^a f(x)dx = 0$

(b) $\int_a^b f(x)dx = -\int_b^a f(x)dx$

(c) $\int x^r dx = \frac{1}{r+1}x^{r+1} + C$

(d) $\int_a^b f(x)dx = F(b) - F(a)$

(e) $\int x^n dx = \frac{1}{n+1}x^{n+1} + C, n \neq -1$

(f) $\int \frac{1}{x} dx = \ln|x| + C$

(g) $\int e^x dx = e^x + C$

(h) $\int_a^b f(g(x))g'(x) = \int_{g(a)}^{g(b)} f(u)du$, where $u = g(x)$

(i) Integration by parts (分部积分): $\int_a^b udv = uv|_a^b - \int_a^b vdu$

Integrals: Exercises

- $\int_0^5 x^2 e^{-x} dx$
- $\int x \ln x$
- $\int_0^t \frac{2}{1000^2} x e^{-(x/1000)^2}$
- $\int_{-\infty}^{\infty} \frac{1}{2} e^{-|y|+ty}$

Taylor's Expansion (泰勒展开)

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$$

Exercise

$$1 + \frac{1}{2!} + \frac{1}{3!} + \ldots$$

Multivariate Calculus

- $f_x(x,y) = \frac{\partial}{\partial x} f(x,y) = \frac{\partial f}{\partial x}$
- $\int_a^b \left[\int_c^d f(x,y) dx \right] dy = \int_c^d \left[\int_a^b f(x,y) dy \right] dx$

Next Section ...

Linear vector space (线性向量空间)

A linear vector space $\mathcal{X}$ is a collection of elements satisfying the following properties:

- Rule of addition (加法律): $\forall x, y, z \in \mathcal{X}$,
  1. $x + y \in \mathcal{X}$
  2. $x + y = y + x$
  3. $(x + y) + z = x + (y + z)$
  4. $\exists 0 \in \mathcal{X}$, such that $x + 0 = x$
  5. $\forall x \in \mathcal{X}, \exists - x \in \mathcal{X}$ such that $x + (-x) = 0$
- Rule of multiplication (乘法律): $\forall x, y \in \mathcal{X}$ and $a, b \in R$,
  1. $ax \in \mathcal{X}$
  2. $a(bx) = (ab)x$
  3. $1x = x, 0x = 0$
  4. $a(x + y) = ax + ay$

Example: $\mathbb{R}^n$

The n-dimensional Euclidean $\mathbb{R}^n$, is a linear vector space.

Inner product (向量内积)

An **inner product** is a mapping: $\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$.
The inner product between any $x, y \in \mathcal{X}$ is denoted by $\langle x, y \rangle$ and it satisfies the following properties for all $x, y, z \in \mathcal{X}$:

(1) $\langle x, y \rangle = \langle y, x \rangle$

(2) $\langle ax, y \rangle = a\langle x, y \rangle$ for all scalars a

(3) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$

(4) $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \Rightarrow x = 0$

A space $\mathcal{X}$ equipped with an inner product is called an **inner product space**.

Orthongonal Vectors (正交向量)

$\mathbf{x}$ and $\mathbf{y}$ are **orthogonal vectors** if:

$$\langle \mathbf{x}, \mathbf{y} \rangle = 0$$

Let $\mathcal{X} = \mathbb{R}^n$, then

$$\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^T \mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

Norms (向量范数)

The inner product induces the defintion of ℓ_2-norm:

$$\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$

here the norm measure the size (length) of **x**.
The inner product can be written into the following form with norms:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\| \|\mathbf{y}\| \cos \theta$$

where θ is the angle between vectors **x** and **y**.
The general ℓ_p-norm for **x** is:

$$\|\mathbf{x}\|_p = (\sum_i x_i^p)^{1/p}, p = 0, 1, 2, \ldots, \infty$$

We have ℓ_0 and ℓ_1 norms:

$$\|\mathbf{x}\|_0 = \sum_i I(x_i \neq 0), \|\mathbf{x}\|_1 = \sum_i |x_i|$$

Cauchy-Schwartz inequality (柯西-施瓦茨不等式)

$$\langle x, y \rangle \leq \|x\| \|y\|$$

Q: When does the equation hold?

Triangle inequality (三角不等式)

$$\|x - y\| \leq \|x\| + \|y\|$$

Q: When does the equation hold?

Linearly independent (线性无关)

For a set of vectors

$$x_1, x_2, \ldots, x_p \in \mathcal{X},$$

if there exists a set of scalars $a_1, a_2, \ldots, a_p \in \mathbb{R}$ such that not all $a_i = 0$ and

$$\sum_{i=1}^{p} a_i x_i = 0$$

we say that $x_1, x_2, \ldots, x_p$ are **linearly dependent** （线性相关**).**
If equation only holds in the case $a_1 = a_2 = \ldots = a_p = 0$, then we say that the vectors are **linearly independent (线性无关).**

Basis (基)

A set of vectors $\{\phi_i\}(i = 1, \ldots, n)$ is a **basis (基)** for $\mathcal{X}$ if an arbitrary vector $x \in \mathcal{X}$ can be expressed as the linear combination of $\{\phi_i\}(i = 1, \ldots, n)$. That is, there exists a set of scalars $\{\theta_i\}(i = 1, \ldots, n)$, such that

$$x = \sum_{i=1}^{n} \theta_i \phi_i$$

Orthonormal basis (正交基)

The bases $\{\phi_i\}_{i=1}^{n}$ are orthonormal if

$$\phi_i^T \phi_j = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$$

Orthobasis of Hilbert space (希尔伯特空间的正交基)

Every $x \in \mathcal{H}$ can be represented in terms of an orthonormal basis $\{\phi_i\}_{i \geq 1}$ (or "orthobasis" for short) according to:

$$x = \sum_{i \geq 1} \langle x, \phi_i \rangle \phi_i$$

This is easy to see as follows. Suppose x has a representation $\sum_i \theta_i \phi_i$. Then

$$\theta_i = \langle x, \phi_i \rangle$$

Example: Orthonormal basis for $\mathbb{R}^n$

$$\phi_k = [0, \cdots, 1, \cdots, 0]^{-1}$$

where

$$\phi_{k,i} = \begin{cases} 0, & i \neq k \\ 1, & i = k \end{cases}$$

Subspace (子空间)

Consider a set of vectors $\{x_i\}_{i=1}^{p} \in \mathcal{X}$. The **span** of these vectors is the set of all vectors $x \in \mathcal{X}$ that can be generated from linear combinations of the set

$$\mathrm{span}(\{xi\}_{i=1}^{p}) := \left\{ y : y = \sum_{i=1}^{p} a_i x_i, a_1, \ldots, a_p \in \mathbb{R} \right\}$$

This set is also called a subspace of $\mathcal{X}$.
A subset $\mathcal{M} \subset \mathcal{X}$ is a subspace if $x, y \in \mathcal{M}$, we have

$$ax + by \subset \mathcal{M}$$

注：如果$\phi_1, \cdots, \phi_p$是子空间$\mathcal{M} \subset \mathbb{R}^n$的一组正交基，则该子空间中的任意向量$x \in \mathcal{M}$可以写成：

$$x = \sum_{i=1}^{p} \theta_i \phi_i$$

这样虽然$x \in \mathbb{R}^n$，但由于其是$\mathcal{M}$中的向量，所以可以写成p个自由参数的线性组合，也就是说其自由度为p。

Orthogonal projection (正交投影)

Let $\mathcal{H}$ be a **Hilbert space** and let $\mathcal{M} \subset \mathcal{H}$ be a subspace. Every $x \in \mathcal{H}$ can be written as

$$x = y + z$$

where $y \in \mathcal{M}$ and $z \perp \mathcal{M}$, which is shorthand for z orthogonal to $\mathcal{M}$; that is

$$\forall v \in \mathcal{M}, \langle v, z \rangle = 0$$

The vector y is the optimal approximation to x in terms of vectors in M in the following sense:

$$y = \text{argmin}_{v \in \mathcal{M}} \|x - v\|$$

The vector y is called the **orthogonal projection** of x onto $\mathcal{M}$.

Orthogonal subspace projection (正交子空间投影)

Let $\mathcal{M} \subset \mathcal{H}$ and let $\{\phi_i\}_{i=1}^r$ be an orthobasis for $\mathcal{M}$. For any $x \in \mathcal{H}$, the projection of x onto $\mathcal{M}$ is given by

$$y = \sum_{i=1}^{r} \langle \phi_i, x \rangle \phi_i$$

and this projection can be viewed as a sort of filter that **removes all components of the signal x that are orthogonal to $\mathcal{M}$.**

Example

Let $\mathcal{H} = \mathbb{R}^2$. Consider the canonical coordinate system $\phi_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\phi_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Let $\mathcal{M}$ be the subspace spanned by ϕ_1. The projection of any $x = [x_1 \; x_2]^T \in \mathbb{R}^2$ onto $\mathcal{M}$ is

$$\begin{aligned} P_1 x &= \langle x, \phi_1 \rangle \phi_1 \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} x_1 \\ 0 \end{bmatrix} \end{aligned}$$

Orthogonal projections in Euclidean subspaces

More generally suppose we are considering $\mathbb{R}^n$ and we have a orthonormal basis $\{\phi_i\}_{i=1}^r$ for some r-dimensional ($r < n$) subspace $\mathcal{M}$ of $\mathbb{R}^n$. Then the projection matrix is given by

$$P_{\mathcal{M}} = \sum_{i=1}^{r} \phi_i \phi_i^T$$

Moreover, if $\{\phi_i\}_{i=1}^r$ is a basis for $\mathcal{M}$, but not necessarily orthonormal, then

$$P_{\mathcal{M}} = \Phi(\Phi^T\Phi)^{-1}\Phi^T$$

where $\Phi = [\phi_1, \ldots, \phi_r]$, a matrix whose columns are the basis vectors.

注：这被用在线性回归模型$y = X\beta$的求解上，其最小二乘解析解就是y到X张成的p-维子空间的正交投影：

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

Eigendecomposition of a symmetric matrix (对称阵的特征分解)

Let $C \in \mathbb{R}^{n\times n}$ is a real, symmetric matrix ($C^T = C$). $v \in \mathbb{R}^n$ is the **eigenvector (特征向量)** of C such that:

$$Cv = \lambda v$$

where λ is the eigenvalue (特征值) of C corresponding to v. There are n orthonormal eigenvectors for C such that

$$\langle v_i, v_j \rangle = \delta_{ij}$$

Let $V = [v_1, \dots, v_n]$, then

$$C = V \Lambda V^T$$

where $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$.

Singular value decomposition (SVD, 奇异值分解)

The SVD of an $n \times p$ matrix H is written as

$$H = \underbrace{U}_{n\times p} \underbrace{\Sigma}_{p\times p} \underbrace{V^T}_{p\times p}$$

- $U = [u_1, \cdots, u_p]$ where $\{u_i\}_{i=1}^p$ are real n-dimensional vectors, and called the **left singular vectors** of H. $U^T U = I_p$.
- $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_p), \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. And $\{\sigma_i\}_{i=1}^p$ are called the **singular values** of H.
- $V = [v_1, \cdots, v_p]$ where $\{v_i\}_{i=1}^p$ are p-dimensional vectors, and called the **right singular vectors** of H. $V^T V = I_p$.

SVD

Also note that:

$$
\begin{aligned}
H^T H &= (U\Sigma V^T)^T U\Sigma V^T \\
&= V\Sigma U^T U\Sigma V^T \\
&= V\Sigma^2 V^T \\
HH^T &= U\Sigma^2 U^T
\end{aligned}
$$

Therefore,

- $\{\sigma_1^2, \cdots, \sigma_p^2\}$ are the eigenvalues of $H^T H$ and $\{v_1, \cdots, v_p\}$ are the corresponding eigenvectors.
- $\{\sigma_1^2, \cdots, \sigma_p^2\}$ are the p-first eigenvalues of HH^T (the remaining $n-p$ eigenvalues are all zeros) and $\{u_1, \cdots, u_p\}$ are the associated eigenvectors.

Application of SVD

Say we want to solve an **over-determined** linear equations:

$$\underbrace{y}_{n\times 1} = \underbrace{X}_{n\times p} \underbrace{\beta}_{p\times 1}$$

- If $n = p$ and $X = U\Sigma V^T$ with $\sigma_1 \geq \cdots \geq \sigma_p > 0$, we say X is **square and non-singular**, $\beta = X^{-1}y$
- If $n > p$ and $X = U\Sigma V^T$ with $\sigma_1 \geq \cdots \geq \sigma_p > 0$, we say X is **non-square and non-singular**, $\beta = (X^TX)^{-1}X^Ty$. This is called the least squares solution to the over-determined linear equations.
- When $n < p$, this is an **under-determined** linear equations, and can be solved using **penalized regression**.

Assignment 1

1. Can you extend the derivatives to vector/matrix form, say, $\frac{dx^T y}{dx}$, $\frac{dx^T y}{dx^T}$, $\frac{dx^T Ay}{dx}$?
2. What is least squares fitting of $\mathbf{X}\beta = \mathbf{y}$? Can you use the above matrix derivatives to reach the normal equation?
3. What is QR-decomposition and Cholesky decomposition? Can you give some comments on the application of the two decomposition techniques?
4. `base::qr()` and `chol()` can be used to compute the two kinds of decompositions. Give an example to illustrate the usage of the decompositions.
5. What kinds of matrices are positive-definite? positive semi-definite?

Next Section ...

Mathematical Optimization

- Optimization uses a rigorous mathematical model to determine the most efficient solution to a described problem.
- You should identify an objective function
  - Objective is a quantitative measure of the performance
  - Objective is usually a single number

Classification of optimization problems

Optimization problem can be constrained or unconstrained.

Common groups

- Linear programming (LP)
  - Objective function and constraints are both linear
  - $\min_{\mathbf{x}} \mathbf{c}^T\mathbf{x}$ s.t. $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ and $\mathbf{x} \geq 0$
- Quadratic programming (QP)
  - Objective function is quadratic and constraints are linear
  - $\min_{\mathbf{x}} \mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x}$ s.t. $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ and $\mathbf{x} \geq 0$
- Nonlinear programming (NLP)
  - Objective function or at least one constraint is nonlinear

Nonlinear optimization

Dimensionality	One-dimensional	Multi-dimensional		
Category	Non-gradient based	Gradient-based	Hessian-based	Non-gradient based
Algorithms	Golden Section Search	Gradient descent	Newton/Quasi-Newton (L-BFGS, BFGS)	Nelder-Mead

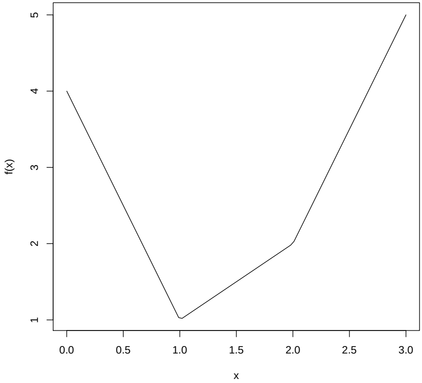
One-dimensional nonlinear programming

- Golden section search
- Basic steps:
  1. Golden ratio: $\phi = (\sqrt{5} - 1)/2 = .618$
  2. Pick an interval $[a, b]$ containing the optimum
  3. Evaluate $f(x_1)$ at $x_1 = a + (1 - \phi)(b - a)$ and compare with $f(x_2)$ at $x_2 = a + \phi(b - a)$
  4. if $f(x_1) < f(x_2)$, continue the search in the interval $[a, x_1]$, else $[x_2, b]$
- R command `stats::optimize()`

```
optimize(f=, interval=, ...,
        tol = .Machine$double.eps^0.25)
```

Nondifferentiable function

- $f(x) = |x - 2| + 2|x - 1|$
- How to solve it?



Newton-Raphson

- Newton method is often used to find the zeros of a function
- Minima fulfill the conditions $f'(x^*) = 0$ and $f''(x^*) > 0$, so Newton can be used to find the zeros of the first derivative
- Basic steps
  1. Approximate the function at the starting point with a linear tangent (e.g., second-order Taylor expansion $t(x) \approx f'(x_0) + (x - x_0)f''(x_0))$
  2. Find the intersect $t(x_i) = 0$ as an approximation to $f'(x^*) = 0$
  3. Use the intersect as the new starting point
  4. Finally, the algorithm $x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$ is repeated until $f'(x_n)$ is close enough to 0.

Hessian-based: BFGS and L-BFGS-B

- Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm builds on the idea of Newton method to take gradient information into account
- Gradient information comes from an approximation of the Hessian matrix
- No guaranteed conversion; especially problematic if Taylor expansion does not fit well
- L-BFGS-B stands for limited-memory-BFGS-box:
  - Extension of BFGS
  - Memory-efficient implementation
  - Additional handles box constraints

Himmelblau's function

$$f(x,y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$$

- Himmelblau's function is a popular multi-modal function to benchmark optimization algorithms.
- For equivalent minima are located at $f(-3.78, -3.28) = 0$, $f(-2.80, 3.13) = 0$, $f(3, 2) = 0$, $f(3.58, -1.85) = 0$