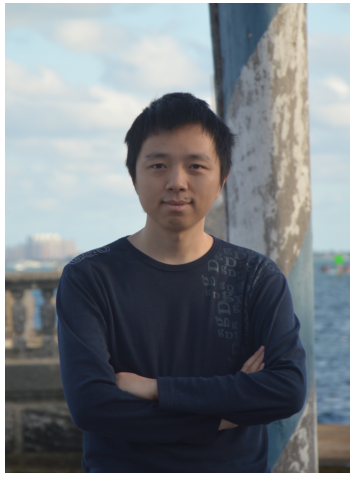
DS-GA 3001: Applied Statistics



Yanjun Han
Instructor

Yuxiao Wen
Section Leader

September 5, 2023

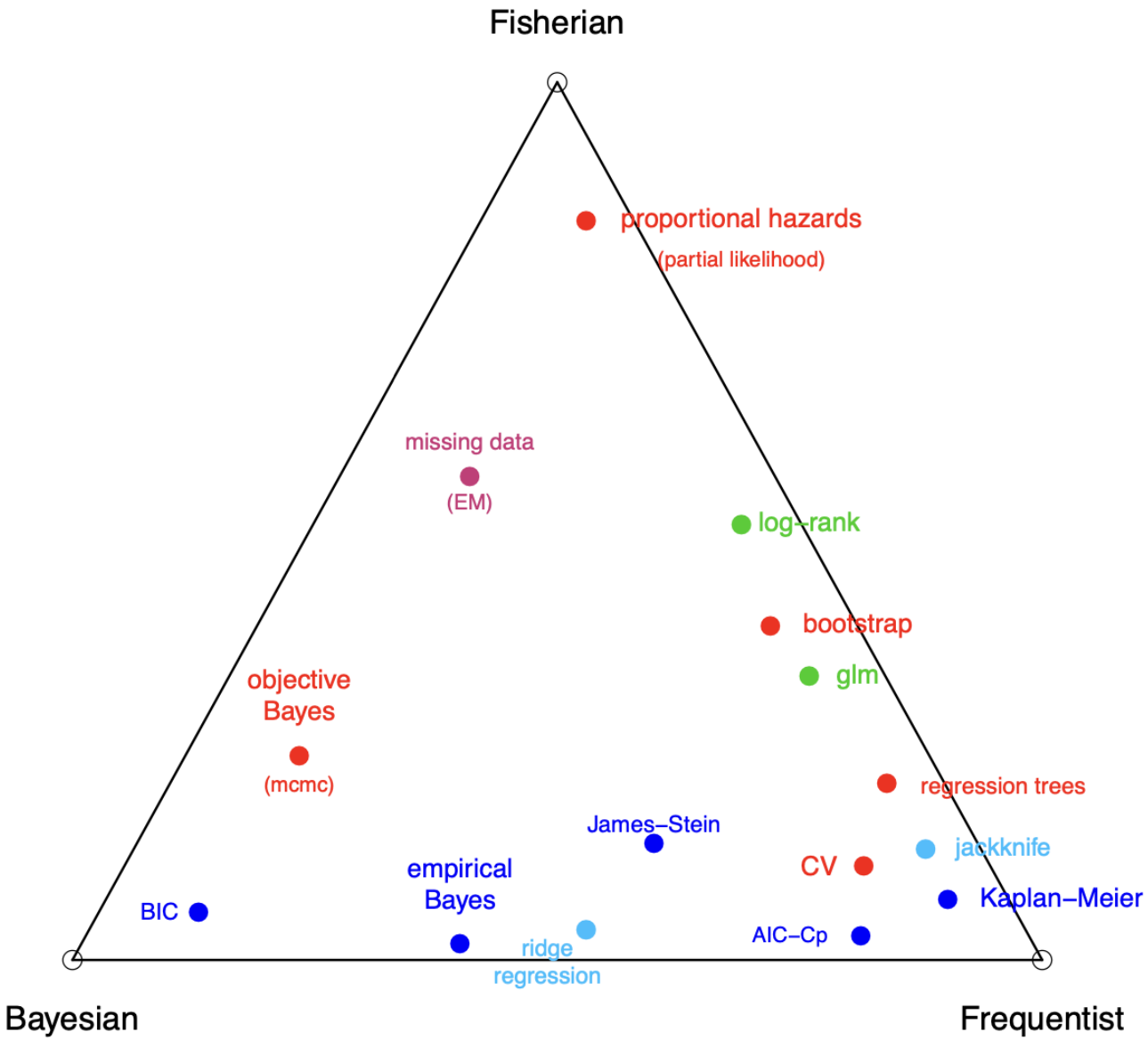
Outline

Course overview

Statistical developments before 1950s

- Frequentist inference
  - bias, variance, correlation
  - linear regression
  - hypothesis testing, Neyman-Pearson
  - p-value, confidence interval
- Bayesian inference
  - prior, posterior
  - Bayes rule
- Fisherian inference
  - likelihood
  - maximum likelihood estimator (MLE)

Important topics after 1950s



What this course is about

- Statistical modeling:
  - parametric model $(x, y) \sim P_\theta$, with finite-dimensional θ;
  - semiparametric model $(x, y) \sim P_{\theta,\eta}$, with nuisance η;
  - nonparametric model $(x, y) \sim P_f$, with infinite-dimensional parameter f;
  - practical burdens: censoring, missing data, confounding, ...
- Statistical tools to solve the above models
- "Applied": focus more on principles/algorithms instead of the analysis

Some questions we want to answer

- Given the nature of a response y, how should I model the noise? (exponential family, semiparametric model)
- In regression, how to choose the link function $\mathbb{E}[y \mid x] = g(\beta^\top x)$? (generalized linear model)
- How to construct a confidence interval for the regressor $\widehat{\beta}$? (bootstrap, deviance residual, likelihood ratio test)
- How to select models with different complexity? (AIC, BIC, Lasso, CV)
- What if we have censored/missing data? (partial/profile likelihood, EM)
- How to set the prior in Bayesian inference? (empirical Bayes)
- How should we deal with the nuisance parameters? (Ichimura's method, double robust estimation)
- What if the regression function is continuous but could be bumpy at times? (local regression, wavelet methods)
- What if the regression function is monotone or convex? (isotonic regression, convex regression, NPMLE)

What this course will cover

- Parametric models: exponential family and GLM ($\sim$ 6 lectures)
  - estimation and inference
  - bootstrap and model selection
  - topic I: censored data and Cox model
  - topic II: missing data and EM algorithm
  - topic III: empirical Bayes
- Semiparametric models ($\sim$ 2 lectures)
  - efficient score and influence
  - topic IV: treatment effect estimation
- Nonparametric models ($\sim$ 4 lectures)
  - linear estimators: KDE, Nadaraya-Watson
  - nonlinear estimators: local regression, splines, wavelet thresholding
  - shape-constrained estimation: isotonic regression, NPMLE

What this course will not cover

- Other linear models: kernel methods and SVM
- Nonlinear models:
  - tree-based models, boosting, random forest
  - neural networks and deep learning
- More advanced topics after this course:
  - probabilistic graphical models
  - Bayesian inference and deep learning
  - causal inference
  - generative models
  - time series analysis

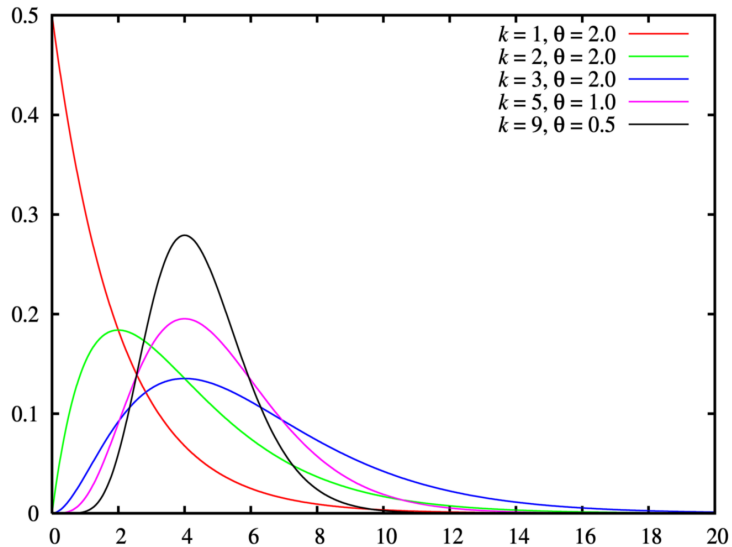
Syllabus

- Lecture 2: Properties of exponential families
- Lecture 3: Parameter estimation and inference
- Lecture 4: Generalized linear model
- Lecture 5: Survival analysis and Cox model
- Lecture 6: Missing data and EM algorithm
- Lecture 7: Empirical Bayes
- (Midterm)
- Lecture 8: Semiparametric models
- Lecture 9: Double robust estimation of average treatment effect
- Lecture 10: Nonparametric density estimation & regression
- Lecture 11: Local polynomial regression and splines
- Lecture 12: Fourier and wavelet methods
- Lecture 13: Isotonic regression & course recap

Lecture 2: exponential family

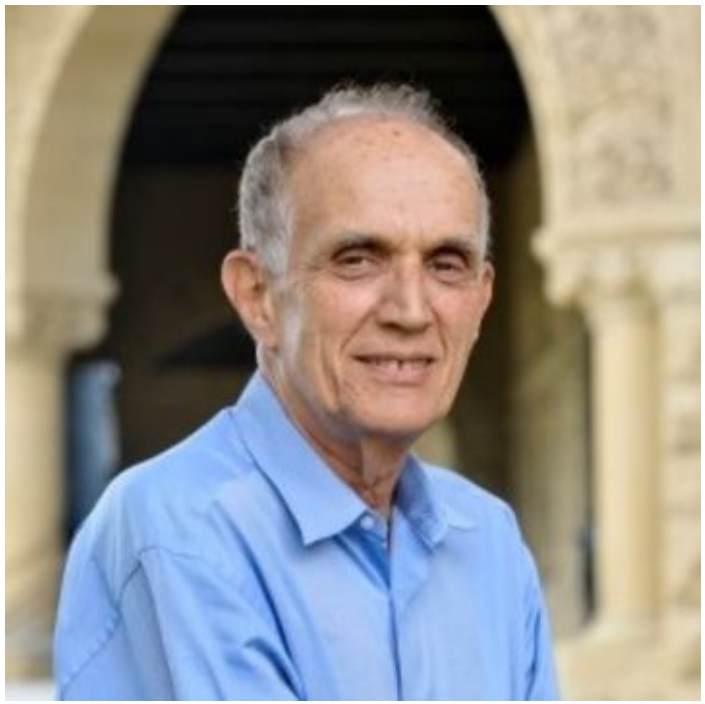
$$p_\theta(y) = \exp(\langle \theta, T(y) \rangle - A(\theta))h(y)$$

- modeling idea
- examples
- properties



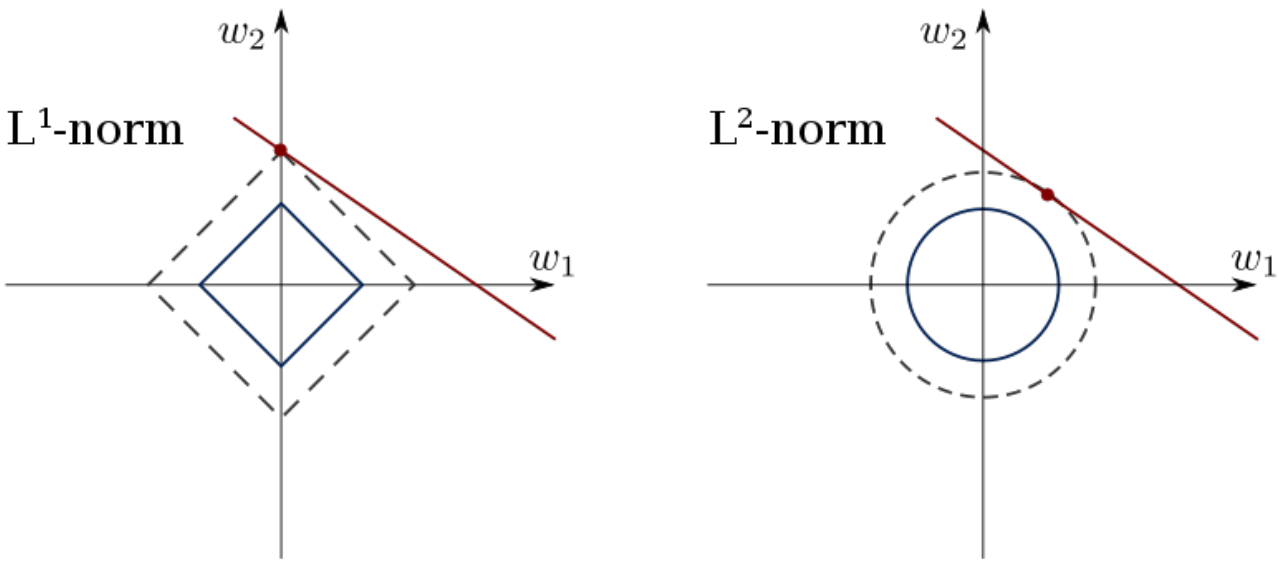
Lecture 3: estimation and inference

- estimator: MLE
- variance estimation: delta method and bootstrap
- inference: Wald test, Rao's score test, likelihood ratio test



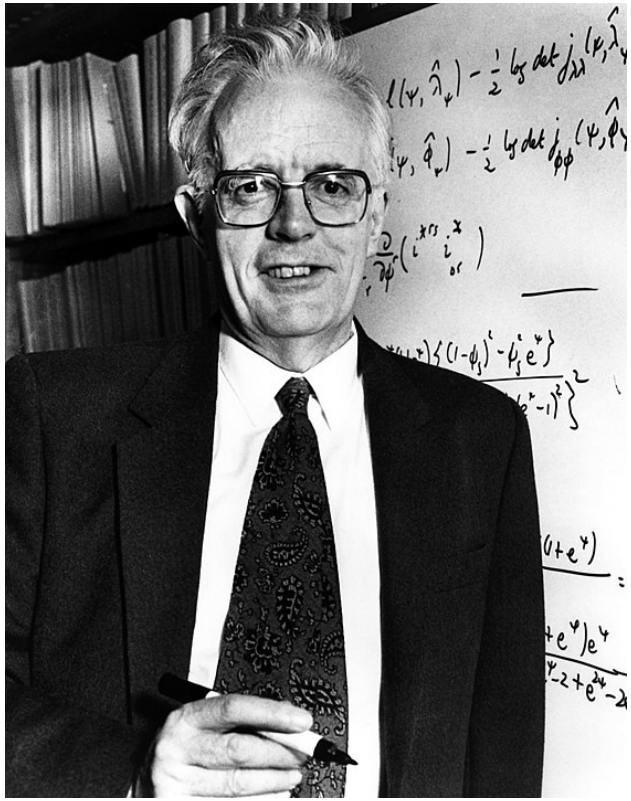
Lecture 4: GLM

- estimation and inference
- model selection: AIC, BIC, and Lasso
- application: Lindsey's method



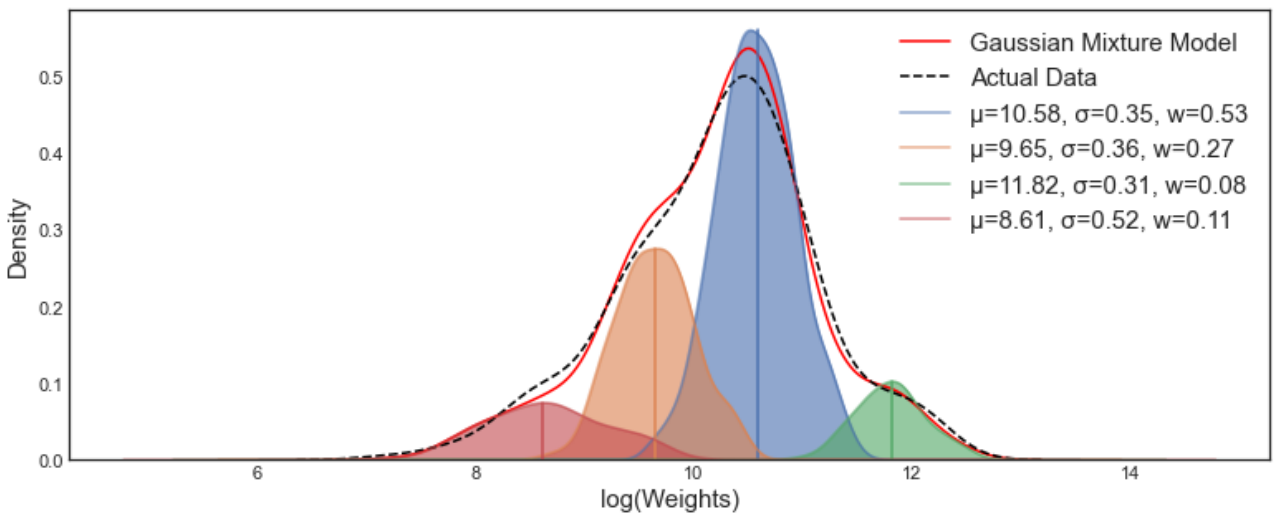
Lecture 5: Cox model

- Kaplan-Meier curve
- Proportional hazards model
- Likelihood modeling:
  - empirical likelihood
  - partial likelihood
  - profile likelihood



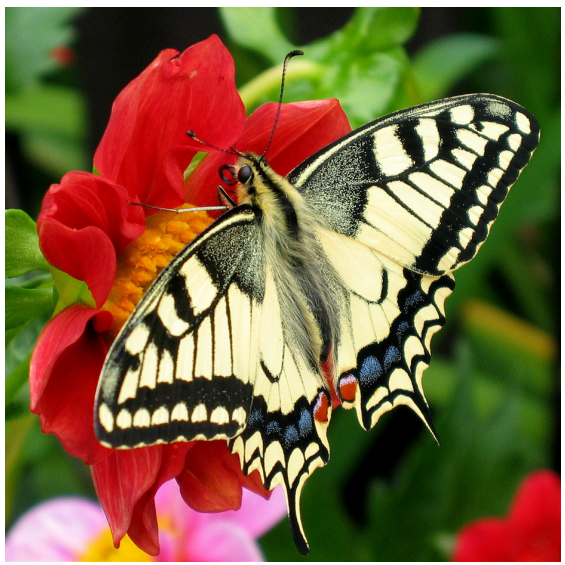
Lecture 6: EM algorithm

- missing data analysis
- convex duality
- extension: variational method in graphical models



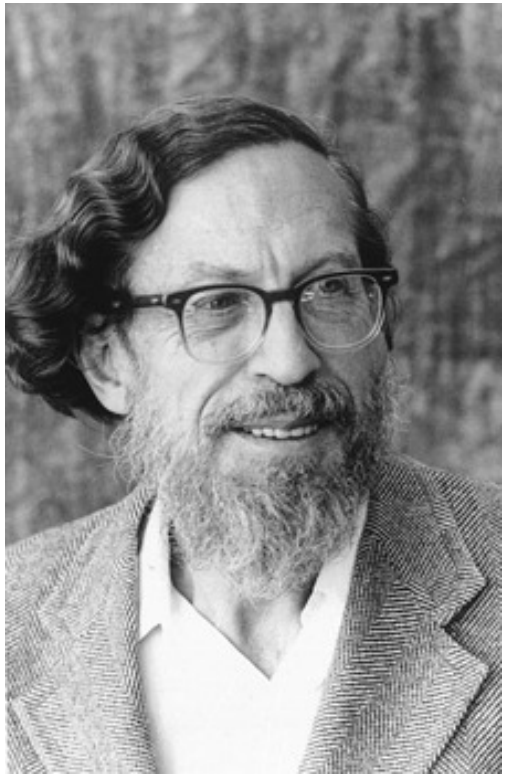
Lecture 7: empirical Bayes

- an old paradigm, but still lots of unknowns
- James-Stein estimator
- Good-Turing estimator
- f-modeling and g-modeling



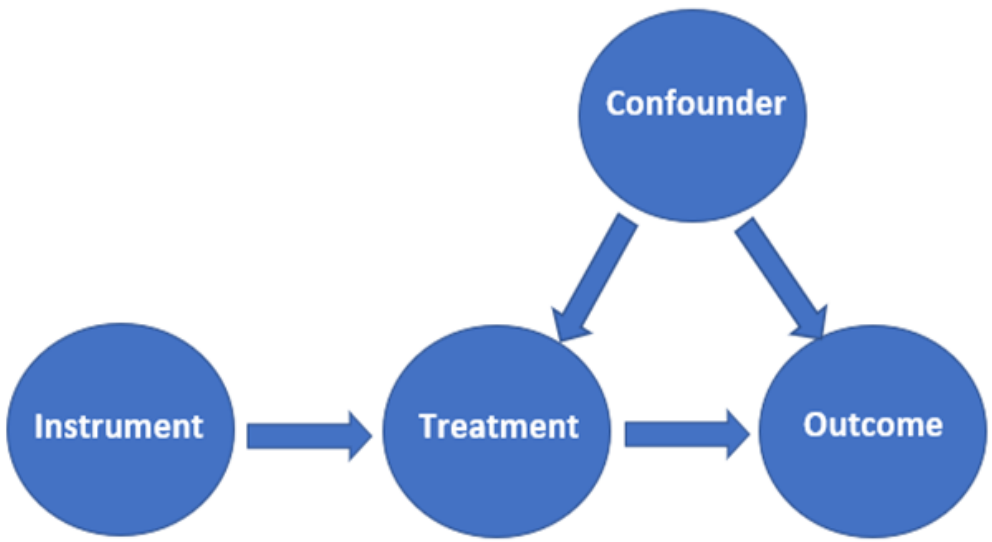
Lecture 8: Semiparametric models

- efficient score and influence
- Ichimura's method
- one-step estimator



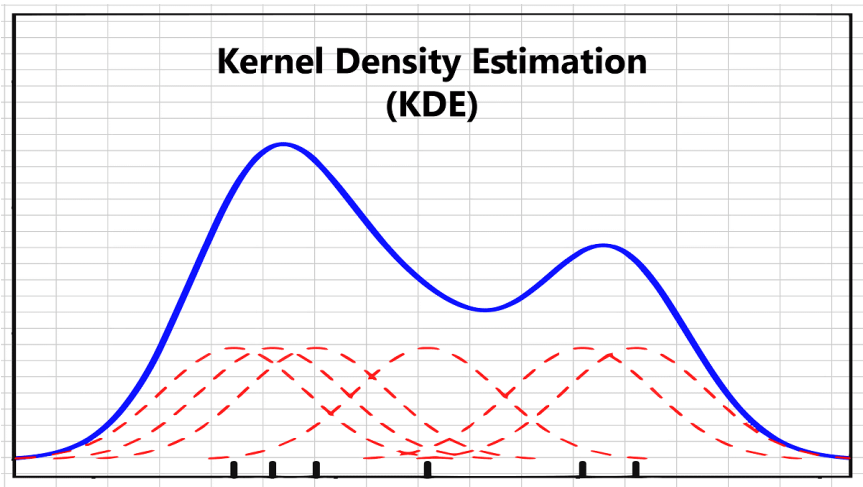
Lecture 9: double robust estimation

- potential outcome model
- double robustness
- extension: double machine learning



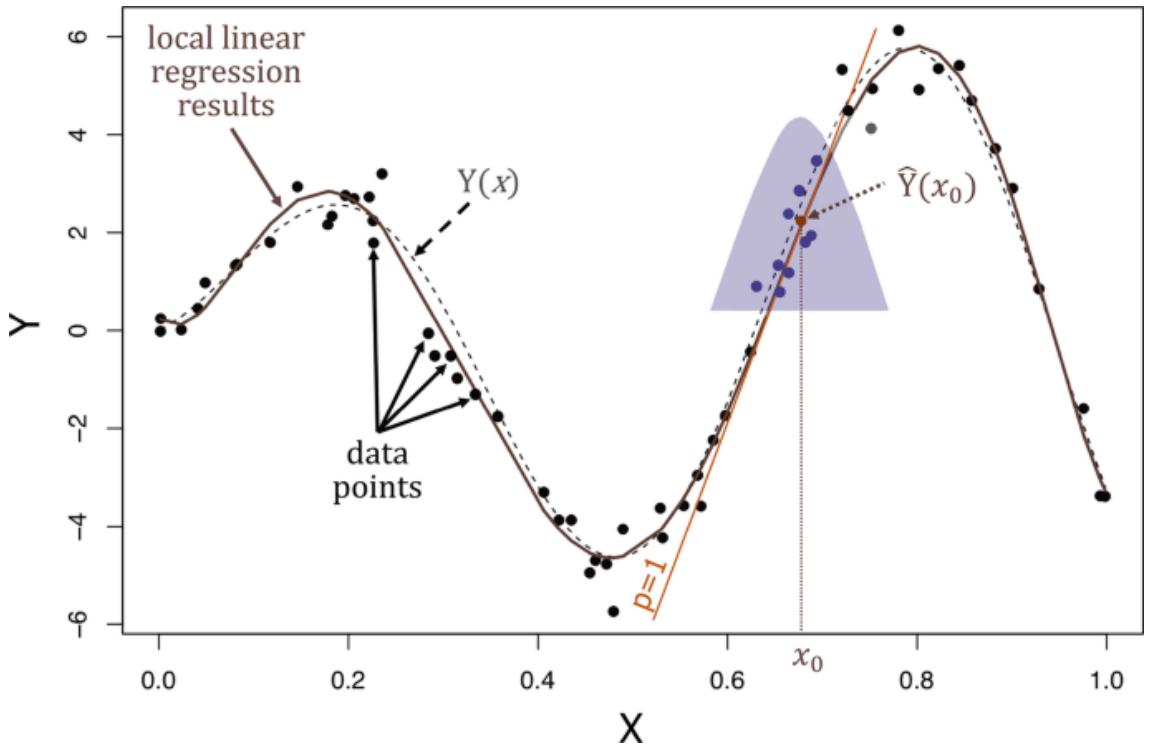
Lecture 10: nonparametric regression

- kernel density estimator
- Nadaraya-Watson estimator
- bias-variance tradeoff



Lecture 11: local regression

- local polynomial regression
- spline regression

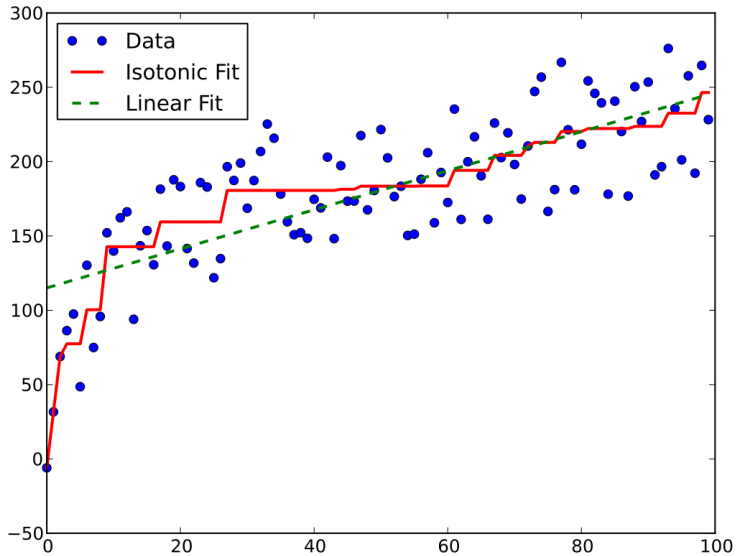


Lecture 12: wavelet methods

- Fourier transform
- wavelet transform
- soft-thresholding and hard-thresholding

Lecture 13: isotonic regression

- shape-constrained regression
- isotonic regression
- nonparametric MLE



Course logistics

Course links

- Course webpage: `https://yanjunhan2021.github.io/courses/applied_stats/index.html`
- Piazza: `https://piazza.com/nyu/fall2023/dsga3001`
- Gradescope: RKXJN2 (entry code)

Meeting times

- Lecture session (Yanjun Han)
  - Tue 4:55 - 6:35 PM at CDS 150 (60 Fifth Ave)
  - regular lectures with handwritten notes
- Lab session (Yuxiao Wen)
  - Fri 2:45 - 3:35 PM at Silver Center 401 (100 Washington Square East)
  - crash course on R, additional examples, homework solutions
- Office hours (starting from 2^{nd} week)
  - Yanjun Han: Thu 4:00 - 5:00pm in CDS 705 and by appointment
  - Yuxiao Wen: Tue 3:00 - 4:00pm in CDS 242 and by appointment

Course notes

- Handwritten notes provided after each lecture

- Reading materials:
  - "Exponential Families in Theory and Practice" by B. Efron
  - "An Introduction to Statistical Learning" by G. James, D. Witten, T. Hastie, and R. Tibshirani
  - "Computer Age Statistical Inference" by B. Efron and T. Hastie
  - "All of Statistics" by L. Wasserman

Homework

- 9 homework assignments in total
- released on Tue, due on Thu 11:59PM of the following week, on Gradescope
- late policy: 2 late HW allowed, each with 3 late days
- the lowest grade will be dropped

Coding assignments

- we will be coding in R
- 1 or 2 coding problems in each HW
- templates in Google colab will be provided, and you only need to fill in a few line of codes and run the script
- Yuxiao will provide a crash course on R in his first two lab sessions

Exams and final grades

- Midterm: Oct 31 in class
- Final: during the final exam week Dec 17-22, details TBD
- Formula for final grades:

$$\text{grade} = 40\% \times \text{HW} + \max\{30\% \times \text{midterm} + 30\% \times \text{final}, 60\% \times \text{final}\}$$

What you can expect from us

- We're here to guide your learning and try to challenge you to engagein the learning process
- We'll do our best to give you tools, feedback, and support to succeed, and please let us know if we can do anything more
- Learning is a never-ending process, so we hope to motivate you to seekout more information on topics we dont have time to cover
- We encourage you to visit us in office hours or to email us for meeting individually. We want to get to know you and support you in this learning experience!

Review of statistical concepts

Statistical estimation

- a collection $(P_\theta)_{\theta \in \Theta}$ of probability distributions is called a statistical model
- an observation is a data point $X \sim P_\theta$ with an unknown $\theta \in \Theta$
- an estimator $T = T(X)$ is a function of x
- the mean squared error (MSE) of T in estimating $g(\theta)$ is

$$\mathrm{MSE}_\theta(T) = \mathbb{E}_{X \sim P_\theta}[(T(X) - g(\theta))^2]$$

- the bias of T is

$$\mathrm{Bias}_\theta(T) = \mathbb{E}_{X \sim P_\theta}[T(X)] - g(\theta)$$

- the variance of T is

$$\mathrm{Var}_\theta(T) = \mathbb{E}_{X \sim P_\theta}[(T(X) - \mathbb{E}_{X' \sim P_\theta}[T(X')])^2]$$

Bias-variance decomposition

$$\mathrm{MSE}_\theta(T) = \mathrm{Bias}_\theta(T)^2 + \mathrm{Var}_\theta(T).$$

MLE and Bayes estimator

- a prior distribution is a probability distribution over θ, i.e. $\theta \sim \pi$
- the posterior distribution is the conditional distribution of θ given X, i.e. $\pi_{\text{post}}(\theta) = \pi(\theta \mid X)$

Bayes rule

The posterior distribution is given by

$$\pi(\theta \mid X) = \frac{\pi(\theta) p_\theta(X)}{p(X)} = \frac{\pi(\theta) p_\theta(X)}{\int \pi(\theta') p_{\theta'}(X) \mathrm{d}\theta'}$$

- the Bayes optimal estimator for $g(\theta)$ with a smallest MSE is the conditional mean $T_{\text{Bayes}}(X) = \mathbb{E}[g(\theta) \mid X] = \int g(\theta) \pi(\theta \mid X) \mathrm{d}\theta$
- the maximum likelihood estimator (MLE) for θ is

$$\widehat{\theta}^{\text{MLE}} = \arg\max_{\theta \in \Theta} p_\theta(X)$$

- the MLE for $g(\theta)$ is the plug-in estimator $g(\widehat{\theta}^{\text{MLE}})$

Simple hypothesis testing

- a simple hypothesis testing problem is to test between

$$H_0 : X \sim P \quad \text{vs} \quad H_1 : X \sim Q$$

- H_0, H_1 are called the null hypothesis and alternative hypothesis, respectively
- a test is a binary function $\Phi(X) \in \{0, 1\}$
- the type-I error is the false positive probability $P(\Phi(X) = 1)$
- the type-II error is the false negative probability $Q(\Phi(X) = 0)$
- (significance) level $=$ type-I error, and power $= 1 -$ type-II error

Neyman-Pearson Lemma

The likelihood ratio test

$$\Phi(X) = \mathbb{1}\left(\frac{Q(X)}{P(X)} \geq \gamma\right)$$

for an appropriate $\gamma > 0$ is the most powerful test among all tests with a given type-I error.

p-value and confidence interval

- a two-sided testing problem is based on $X \sim P_\theta$, test between

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

- a p-value is a function $p = p(X)$ such that $p \sim \text{Unif}([0,1])$ if the null holds
  - consequently, the test $\Phi(X) = \mathbb{1}(p(X) \leq \alpha)$ has significance level α
- a confidence interval with confidence $1 - \alpha$ is an interval $C = C(X)$ such that $\mathbb{P}_\theta(\theta \in C(X)) \geq 1 - \alpha$ for every θ

General recipe for constructing p-values and confidence intervals

- find a function $T = T(X, \theta)$ such that T has a known CDF F when $X \sim P_\theta$;
- the p-value is typically constructed by

$$p = 1 - F(T(X, \theta_0));$$

- the confidence interval is typically constructed by

$$C = \{\theta : T(X, \theta) \in [F^{-1}(\alpha/2), F^{-1}(1 - \alpha/2)]\}.$$

p-value and confidence interval: an example

Let $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with unknown (μ, σ^2), and we are interested in μ

- finding the function T: under true mean μ,

$$T(X_1, \cdots, X_n; \mu) = \frac{\overline{X} - \mu}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \overline{X})^2}} \sim t_{n-1}$$

  follows the Student's t-distribution with $n-1$ degrees of freedom

- p-value for testing $\mu = 0$: compute

$$p = 1 - t_{n-1}(T(X_1, \cdots, X_n; \mu = 0))$$

- confidence interval for μ:

$$\begin{aligned} C &= \{\mu : T(X_1, \cdots, X_n; \mu) \in [t_{n-1}^{-1}(\alpha/2), t_{n-1}^{-1}(1-\alpha/2)]\} \\ &= \left[\overline{X} - \frac{t_{n-1}^{-1}(1-\alpha/2)}{\sqrt{n}}\widehat{\sigma}, \overline{X} + \frac{t_{n-1}^{-1}(1-\alpha/2)}{\sqrt{n}}\widehat{\sigma}\right] \end{aligned}$$

A bit convex analysis

- a set A is convex iff for $a, b \in A$ and $\lambda \in [0,1]$, we have $\lambda a + (1-\lambda)b \in A$
- a function f is convex if $\mathrm{dom}(f)$ is convex, and

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b), \quad \forall a, b \in \mathrm{dom}(f), \lambda \in [0,1]$$

- a function f is concave if $-f$ is convex

Some properties of convex functions

- if f, g are convex and $\alpha \geq 0$, so are $f+g$ and αf;
- if $x \mapsto f(x)$ is convex, then for any matrix A, $y \mapsto f(Ay)$ is also convex;
- if $(f_i)_{i \in I}$ are all convex, so is $f(x) = \max_{i \in I} f_i(x)$;
- a twice continuously differentiable function f is convex iff the Hessian matrix $\nabla^2 f \succeq 0$ is PSD everywhere.