Overview of Statistical Evolution

Yanrong Yang

RSFAS/CBE, Australian National University

26th July 2022

Introduction to this course

- Delivery
  3-hour lecture (Weeks 1 - 12)
  1-hour tutorial class (Weeks 2 - 12)
- Assessments
  3 assignments (NOT Redeemable) + final exam
- Lecturer: Yanrong Yang (yanrong.yang@anu.edu.au)
  Tutor: Yonghe Lu (yonghe.lu@anu.edu.au)
- Lecturer Consultation
  1:00pm - 3:00pm each Friday (Weeks 1 - 12)
  Meeting ID: 239 899 532
  Password: 049569
- Tutor Consultation
  1:00pm - 2:00pm each Monday (Weeks 2 - 12)
  Meeting ID: 895 0151 5248
  Password: 805980

Contents of this week

- Core and Evolution of Statistical Analysis
- Modern Data: Popularity and Complexity of Big Data
- Modern Statistics: Introduction to Advanced Statistical Learning
- Necessity of Statistical Learning: Illustration of Big Data in Practice

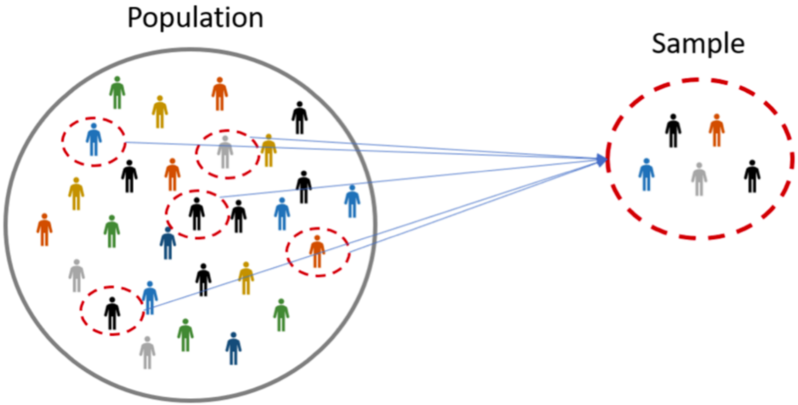
Review on Statistical Analysis

What is Statistical Analysis?

Statistical analysis is the science of learning from experience.

- Aim: Use collected data to infer the population
- Job: Estimation (or Algorithm) and Inference (or Assessment)
- Essential Feature of Statistics: Uncertainty (or Randomness)
- Challenge: model flexibility

Uncertainty



Example 1: Sample Mean

Consider the classical estimator for μ below

$$Sample\ Mean : \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1}$$

Assessment of Sample Mean:

1. Bias: $Bias(\overline{x}) = \mathbb{E}\left(\overline{x}\right) - \mu = 0.$
2. Variance: $Var(\overline{x}) = \mathbb{E}\left(\overline{x} - \mathbb{E}\left(\overline{x}\right)\right)^2 = \frac{\sigma^2}{n}.$

For example, $\mu = 1$, $\sigma^2 = 1$ and $n = 100$

Exp.	1	2	3	4	5	6	7	8
$\overline{x}$	1.123	0.927	0.889	0.935	1.016	1.043	0.962	0.904

Example 1: Asymptotic Theory

Large Sample Theory (or Asymptotic Theory) provides beautiful results for Statistics.

1. Law of Large Numbers (LLN):

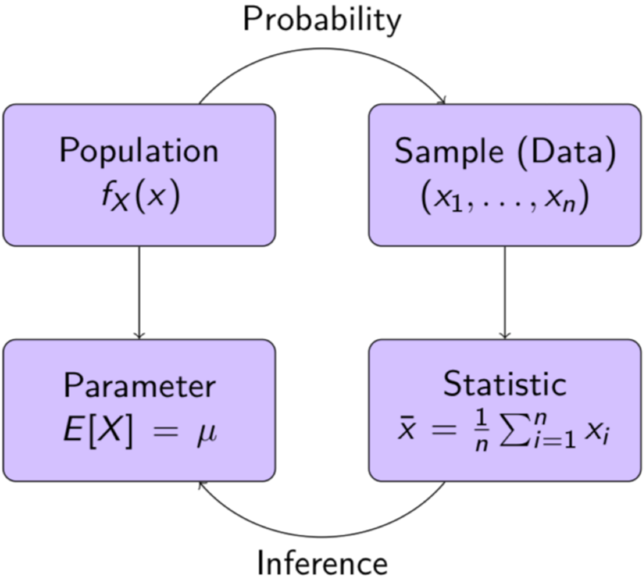
$$\overline{x} \xrightarrow{i.p.} \mu, \; as \; n \to \infty. \tag{2}$$

2. Central Limit Theorem (CLT):

$$\frac{\sqrt{n}\,(\overline{x} - \mu)}{\sigma} \xrightarrow{i.d.} N(0,1), \quad as \; n \to \infty. \tag{3}$$

CLT can also help to do hypothesis test for some hypothesis on the parameter μ.

Formal Setup for Statistical Inference

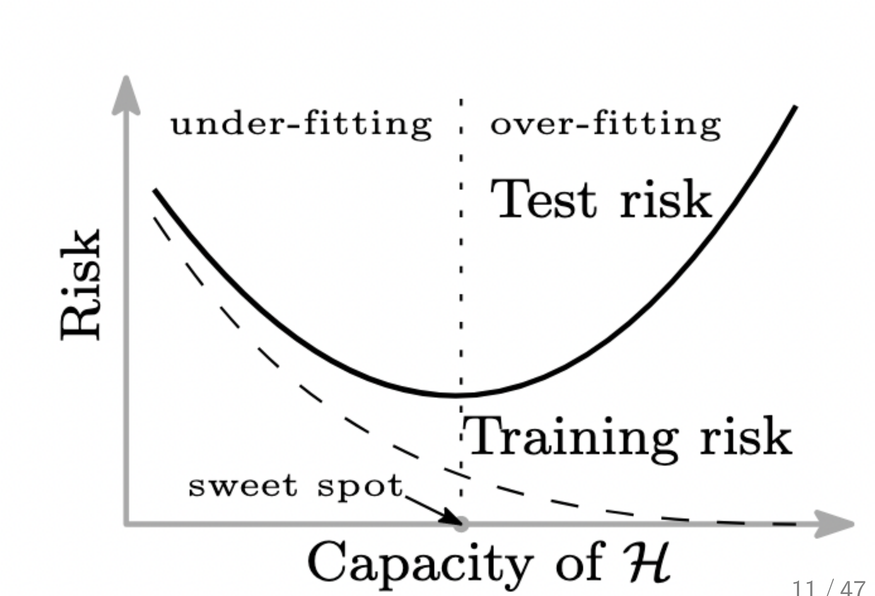


Model Flexibility

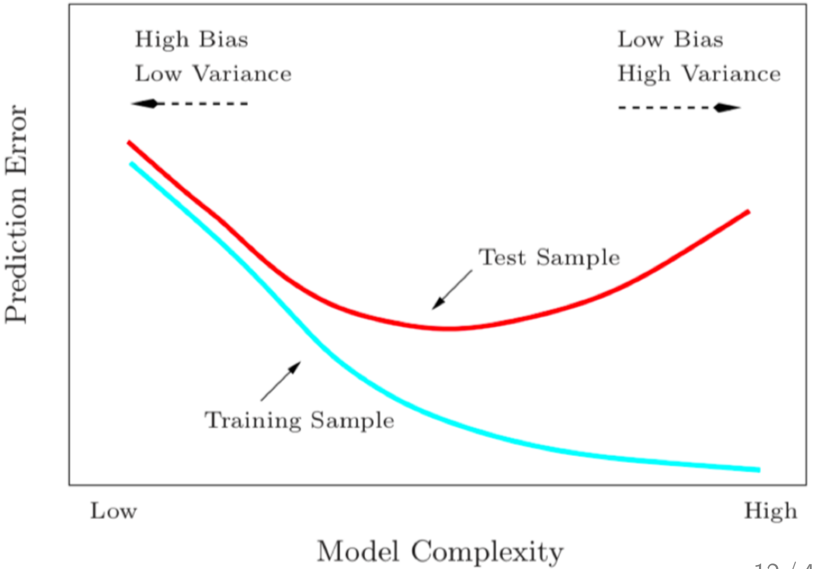
Determining model flexibility (or model selection) is the key challenge for statistical analysis.

- Test mean square error (MSE) is an important criterion to assess statistical methods.
- Small test MSE requires both small bias and small variance.
- Large flexibility always results in small bias but large variance.
- Appropriate flexibility comes from trade-off between bias and variance.

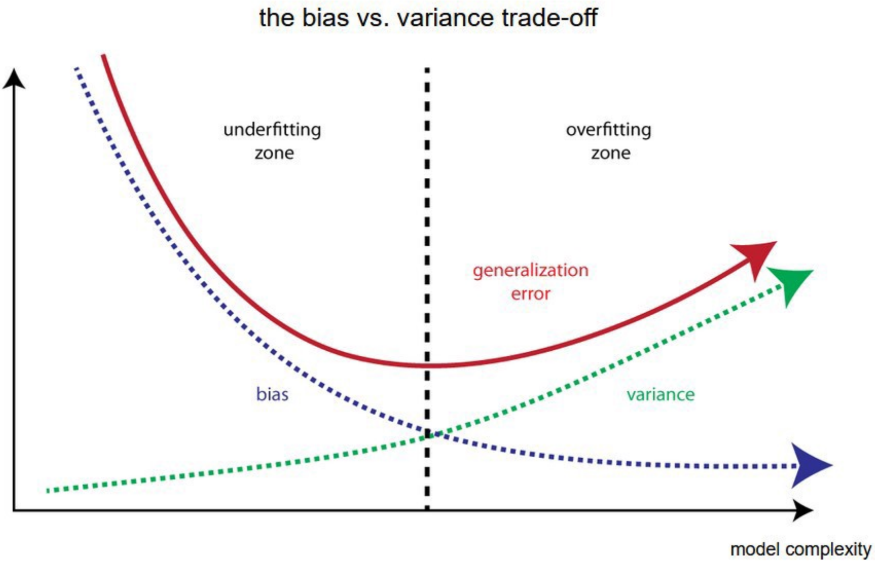
Tradeoff between Bias and Variance (1)



Tradeoff between Bias and Variance (2)



Tradeoff between Bias and Variance (3)



Example 2: Study on Kidney Function

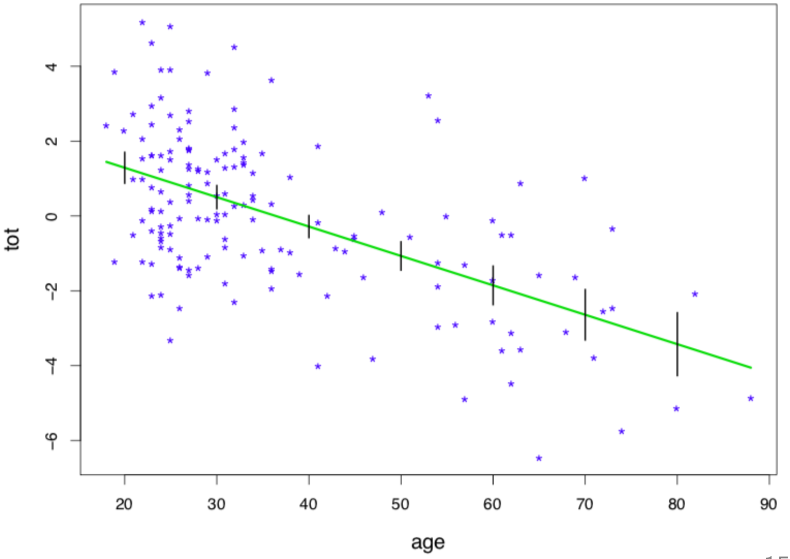
Data

- Data points (x_i, y_i) are observed for $n = 157$ healthy volunteers.
- x_i: the i-th volunteer's age in years.
- y_i: a composite measure "tot" of overall function.
- The rate of Kidney function decline (with age) is an important question in kidney transplantation.

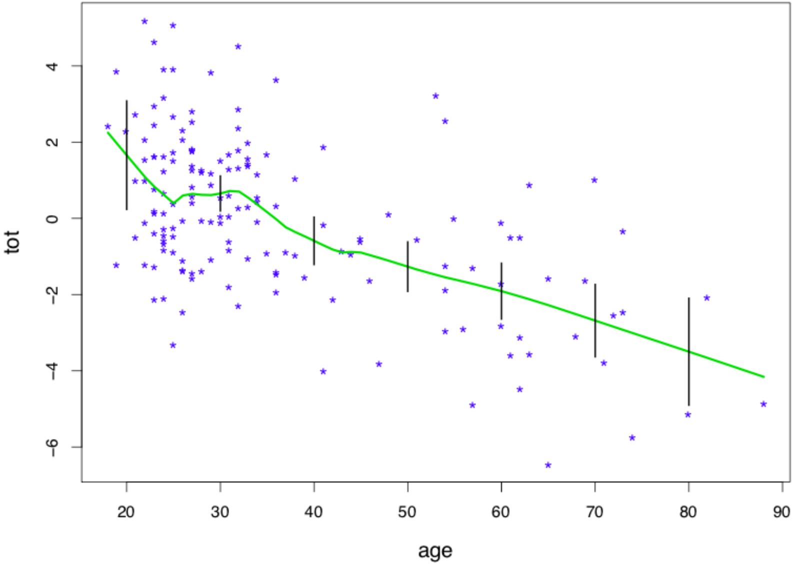
Methods

- Linear Regression: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.
- Quadratic Regression: $y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \eta_i$.

Example 2: Linear Regression



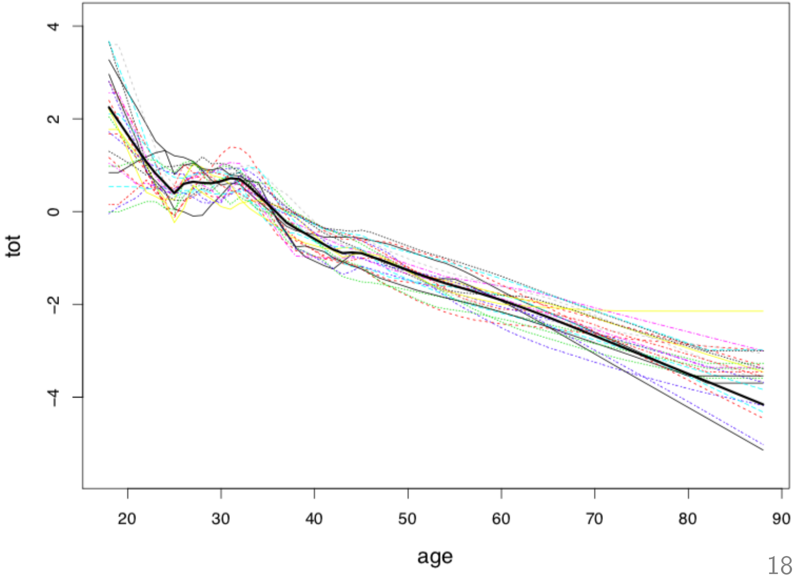
Example 2: Local Polynomial Regression



Example 2: Comparison

age	20	30	40	50	60	70	80
1. linear regression	1.29	.50	−.28	−1.07	−1.86	−2.64	−3.43
2. std error	.21	.15	.15	.19	.26	.34	.42
3. lowess	1.66	.65	−.59	−1.27	−1.91	−2.68	−3.50
4. bootstrap std error	.71	.23	.31	.32	.37	.47	.70

Example 2: Bootstrap Replications



Introduction to Advanced Statistical Learning

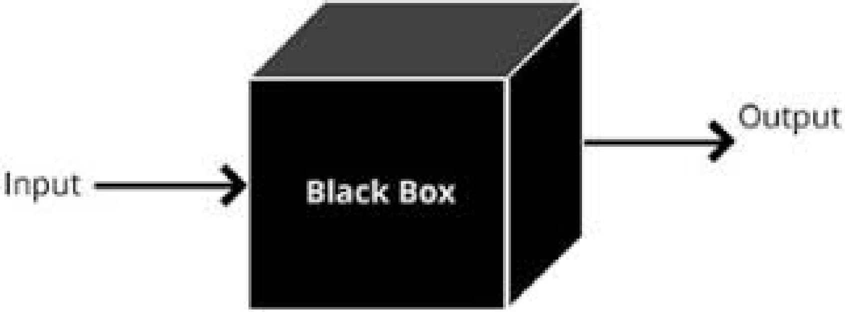
What is Advanced Statistical Learning?

Statistical Learning indicates modern statistical analysis, which is based on intensive computation.

- Big Data: Curse of Dimensionality and Heterogeneity
- New Methodologies/Algorithms (Machine Learning): deep neural networks, adaboosting, support vector machines, ...
- Inference: statisticians try to locate the new methodology within the framework of statistical theory.

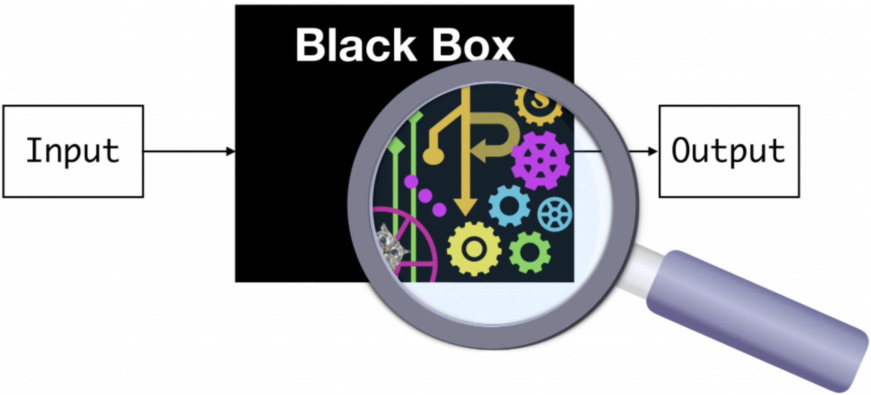
This is a healthy chain of events, good both for the hybrid of the statistics profession and for the further progress of algorithm technology.

Machine Learning



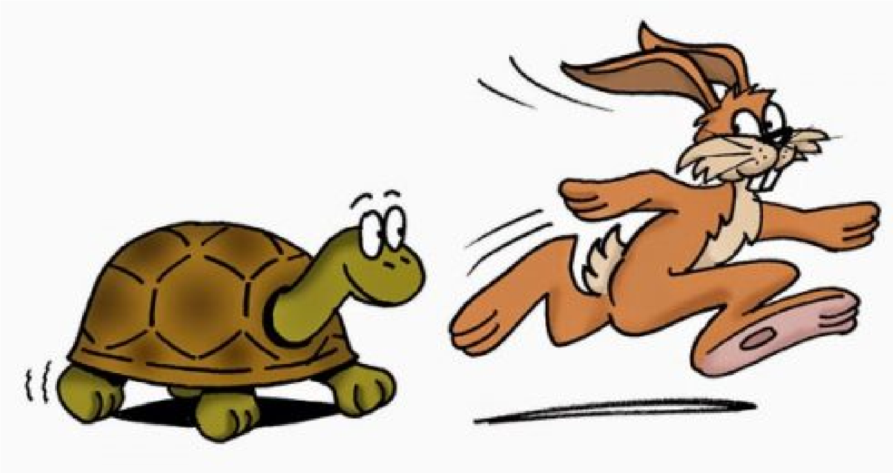
Statistical Learning

Statistical Learning aims to statistical proproties of algorithms in the black box.



Relation between Algorithm and Inference

The inference and algorithm race seems to be like a tortoise-and-hare affair.



History of Statistical Analysis

Evolution of Statistics

- Classical Inference:
  Frequentist Inference, Bayesian Inference, Fisherian Inference
- Early Computer-Aga Methods:
  Shrinkage methods (James-Stein Estimation, Ridge Estimation), Generalized Linear Models, the EM Algorithm, MCMC, the Jacknife and the Bootstrap, ...
- Twenty-First-Century Topics:
  Random Forests, Boosting, Neural Networks, Deep Learning, Kernel Methods, ...

Twenty-First-Century Statistics

- The major challenge of modern statistics includes curse of dimensionality and heterogeneity.
- New findings in inference also appear in modern statistics.

Example 3: High-dimensional Regression

Consider a linear regression model below

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where y_i and $\mathbf{x}_i : p \times 1$ represent dependent variable and independent variables respectively; $\boldsymbol{\beta} : p \times 1$ is vector of unknown parameters; and ε_i is error component. Ordinary Least Squares (OLS) estimator is

$$\widehat{\boldsymbol{\beta}}_{OLS} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{'} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i y_i \right). \tag{2}$$

How is the behaviour of $MSE = \mathbb{E} \left|\left| \widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta} \right|\right|^2$?

Example 3: Challenge

Let us try a simulation on MSE. $\boldsymbol{\beta} = \mathbf{1}_p$, $n = 100$, $\varepsilon_i \sim N(0,1)$, $\mathbf{x}_i \sim N(\mathbf{1}_p, \mathbf{I}_p)$, and MSE in the following table is estimated by average over 1000 simulations.

p	10	20	50	60	80	100	200
MSE	0.983	2.162	7.015	9.391	18.052	468.7	NA

This table shows that the least-square estimator becomes inaccurate as the dimension p increases.

Example 4: High-dimensional Mean Vector

Estimator: **Sample Mean**

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \tag{10}$$

Norm Bias:

$$Norm\ Bias = ||\mathbb{E}\left(\bar{\mathbf{x}}\right) - \boldsymbol{\mu}|| = ||\mathbf{0}_p|| = 0. \tag{11}$$

Norm Variance:

$$Norm\ Variance = \mathbb{E}\left||\bar{\mathbf{x}} - \mathbb{E}\left(\bar{\mathbf{x}}\right)\right||^2 = \frac{1}{n}\sum_{k=1}^{p}\sigma_k^2. \tag{12}$$

In particular, Norm Variance is equal to $\frac{p\sigma^2}{n}$ if $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2$.

Example 4: Challenge

Look at the criterion Mean Squared Error (MSE)

$$\begin{aligned} MSE &= \mathbb{E}\left\|\overline{\mathbf{x}} - \boldsymbol{\mu}\right\|^2 \\ &= \mathbb{E}\left\|\overline{\mathbf{x}} - \mathbb{E}\left(\overline{\mathbf{x}}\right)\right\|^2 + \mathbb{E}\left\|\mathbb{E}\left(\overline{\mathbf{x}}\right) - \boldsymbol{\mu}\right\|^2 \\ &= Norm\ Variance + [Norm\ Bias]^2 = \frac{1}{n}\sum_{i=1}^{p}\sigma_i^2. \end{aligned}$$

1. As p is fixed,

$$MSE \asymp \frac{1}{n} \to 0, \quad as\ n \to \infty.$$

2. As p goes to infinity, for convenience, we assume $\sigma_i^2 = \sigma^2$, $\forall i = 1, 2, \ldots, p$,

$$MSE \asymp \frac{p}{n} \tag{13}$$

Example 4: Results

Let us look at an example: $\boldsymbol{\mu} = \mathbf{1}_p$, $n = 100$ and MSE in the following table is estimated by average over 1000 simulations.

p	10	20	50	100	200	300	400
MSE	0.099	0.197	0.504	0.995	2.010	3.002	3.997

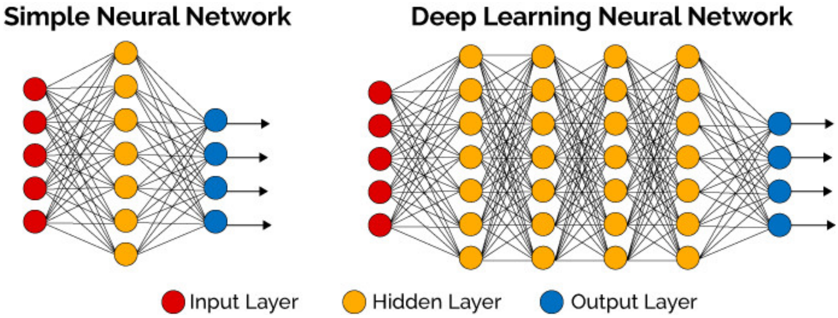
As p increases, the MSE becomes larger. It indicates $\bar{\mathbf{x}}$ as worse estimator when p increases.

Curse of Dimensionality

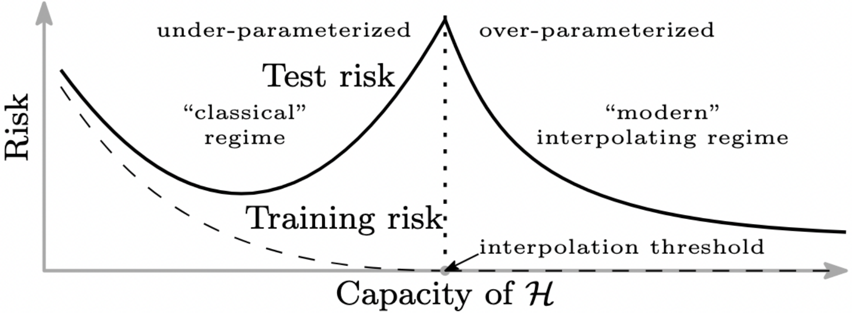
Curse of dimensionality is due to insufficient available information to recover all unknown information. Intuitively and naturally, the solution should be to recover some important information as most as possible, which directs to feature selection. Statistical Learning has shown corresponding techniques on feature selection or feature learning

1. Regularization or Shrinkage Methods: Ridge Estimation and the Lasso;
2. Dimension Reduction: Principal Component Analysis (PCA)and its application Principal Component Regression (PCR).

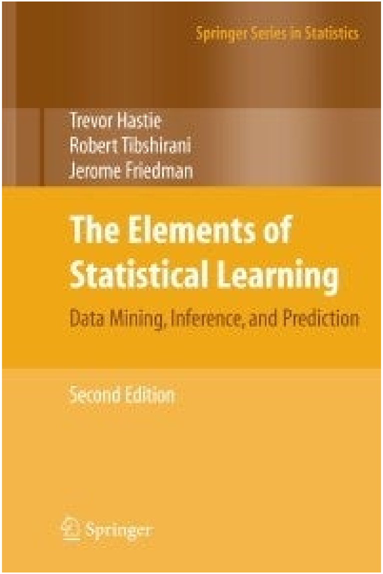
Example 5: Deep Neural Networks



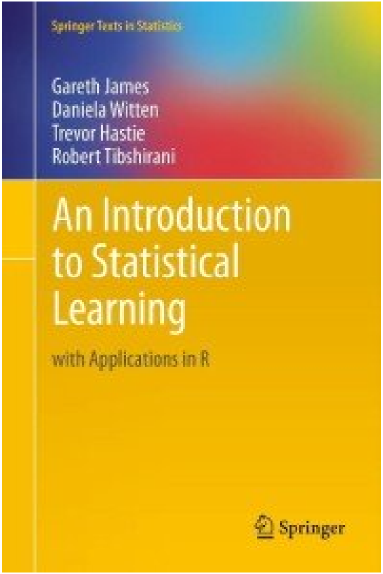
New Finding: Double Descent Phenomenon



Excellent Books on Statistical Learning (1)



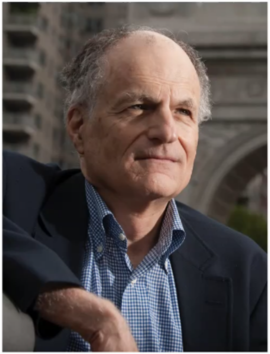
Excellent Books on Statistical Learning (2)



Significance of Statistics

Statistics

Statistics plays a critical role in data-driven AI.



> **"Artificial intelligence is actually statistics,** but it uses a very gorgeous rhetoric. … **all artificial intelligence uses statistics** to solve problems."
>
> — Thomas J. Sargent
> Winner of the 2011 Nobel Prize in Economics

Wharton Tony Cai 4

Big Data in Finance and Actuarial Science

Example 6: Multi-Section Daily Stock Returns

- The data are collected from the Center for Research in Security Prices (CRSP) and include the daily stock returns of 160 companies from 1st January, 2014 to 31st December, 2014, with 252 trading days.
- The 160 stocks are selected from eight different industries according to Fama and French's 48-industry classification, namely, Candy and Soda, Tobacco Products, Apparel, Aircraft, Shipbuilding and Railroad Equipment, Petroleum and Natural Gas, Measuring and Control Equipment, and Shipping Containers, with 20 stocks from each industry.
- The data for the first 126 trading days are treated as the training sample, and the rest are the testing sample. Thus, the training data have the dimensions $n = 126$ and $p = 160$.

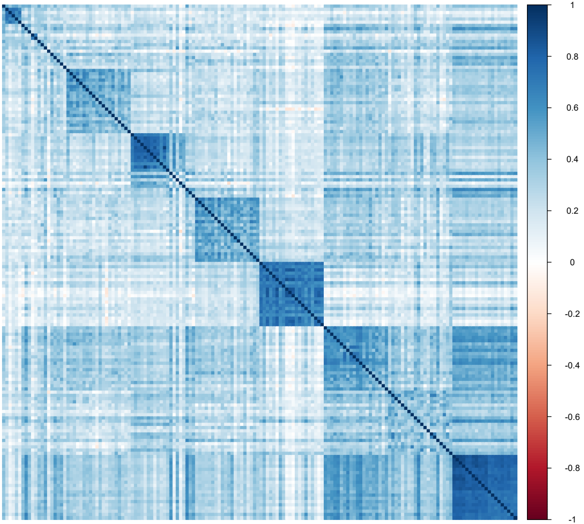
Example 6: Covariance Structure

- For the $p = 160$ stock returns, their observations on day t are involved into a vector $\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots, x_{pt})^\top$. Here $t = 1, 2, \ldots, n = 126$.
- The aim is to extract common factors $\{f_{1t}, f_{2t}, \ldots, f_{rt}\}$ from across the $p = 160$ stock returns.

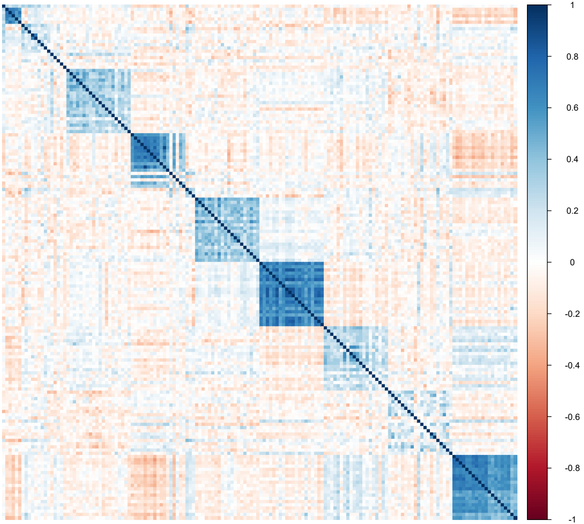
$$x_{it} = \lambda_{i1} f_{1t} + \lambda_{i2} f_{2t} + \cdots + \lambda_{ir} f_{rt} + \varepsilon_{it}$$

- Factor analysis is important for high-dimensional data analysis: (1) interpretation; (2) forecasting.

Correlation plot for the original stock data $\mathbf{x}_t$



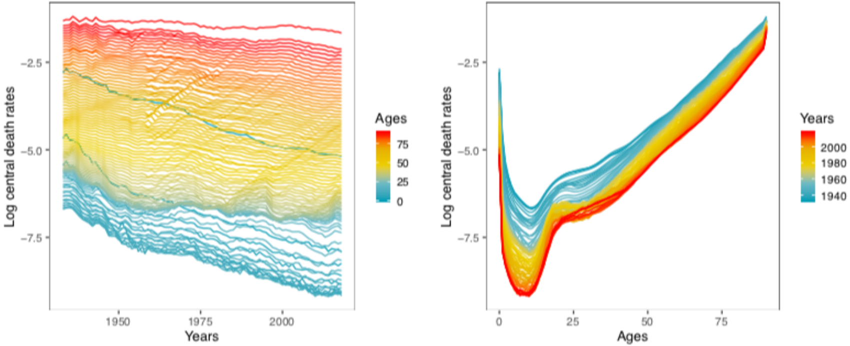
Correlation plot for data after removing factor part



Example 7: Mortality Data

	Historical data					Forecasts		
	1933	1934	1935	…	2018	2019	2020	…
0	−2.792	−2.681	−2.789	…	…	?	?	?
1	−4.661	−4.551	−4.720	…	…	?	?	?
2	−5.437	−5.328	−5.486	…	…	?	?	?
3	−5.775	−5.735	−5.816	…	…	?	?	?
4	−6.038	−6.011	−6.031	…	…	?	?	?
5	−6.227	−6.200	−6.210	…	…	?	?	?
…	…	…	…	…	…	?	?	?
90+	…	…	…	…	…	?	?	?

Example 7: US Mortality



Example 7: Benchmark Literature

Modeling and Forecasting U.S. Mortality

RONALD D. LEE and LAWRENCE R. CARTER*

Time series methods are used to make long-run forecasts, with confidence intervals, of age-specific mortality in the United States from 1990 to 2065. First, the logs of the age-specific death rates are modeled as a linear function of an unobserved period-specific intensity index, with parameters depending on age. This model is fit to the matrix of U.S. death rates, 1933 to 1987, using the singular value decomposition (SVD) method; it accounts for almost all the variance over time in age-specific death rates as a group. Whereas e_0 has risen at a decreasing rate over the century and has decreasing variability, $\mathbf{k}(t)$ declines at a roughly constant rate and has roughly constant variability, facilitating forecasting. $\mathbf{k}(t)$, which indexes the intensity of mortality, is next modeled as a time series (specifically, a random walk with drift) and forecast. The method performs very well on within-sample forecasts, and the forecasts are insensitive to reductions in the length of the base period from 90 to 30 years; some instability appears for base periods of 10 or 20 years, however. Forecasts of age-specific rates are derived from the forecasts of $\mathbf{k}$, and other life table variables are derived and presented. These imply an increase of 10.5 years in life expectancy to 86.05 in 2065 (sexes combined), with a confidence band of plus 3.9 or minus 5.6 years, including uncertainty concerning the estimated trend. Whereas 46% now survive to age 80, by 2065 46% will survive to age 90. Of the gains forecast for person-years lived over the life cycle from now until 2065, 74% will occur at age 65 and over. These life expectancy forecasts are substantially lower than direct time series forecasts of e_0, and have far narrower confidence bands; however, they are substantially higher than the forecasts of the Social Security Administration's Office of the Actuary.

KEY WORDS: Demography; Forecast; Life expectancy; Mortality; Population; Projection.

From 1900 to 1988, life expectancy in the United States rose from 47 to 75 years. If it were to continue to rise at this same linear rate, life expectancy would reach 100 years in 2065, about seventy five years from now. The increase would be welcomed by most of us, but it would come as a nasty surprise to the Social Security Administration, which plans on the more modest life expectancy of 80.5 years predicted by its Office of the Actuary. We scarcely need dwell on the importance of the future course of mortality in our aging society. In contrast to the past, now mortality decline is a powerful cause of population aging.

There are many ways to forecast mortality (Land 1986; Olshansky 1988). The new method we propose here is extrapolative and makes no effort to incorporate knowledge about medical, behavioral, or social influences on mortality change. Its virtues are that it combines a rich yet parsimonious demographic model with statistical time series methods, it is based firmly on persistent long-term historical patterns and trends dating back to 1900, and it provides probabilistic confidence regions for its forecasts. While many methods assume an upper limit to the human life span or rationalize in some other way the deceleration of gains in life expectancy, our method allows age-specific death rates to decline exponentially without limit; the deceleration of

Next we fit the demographic model to U.S. data and evaluate its historical performance. Using standard time series methods, we then forecast the index of mortality and generate associated life table values at five-year intervals. Because we intend our forecasts to be more than illustrative, we present them in some detail and provide information to enable the reader to calculate life table functions and their confidence intervals for each year of the forecast.

1. THE HISTORICAL DATA

Annual age-specific death rates for the entire U.S. population are available for the years 1933 to 1987. For the years 1900 to 1932, these data are available annually only for the death registration states, which form a varying subset of the total U.S. population, and have a cruder age specificity (see Grove and Hetzel 1968, table 51, p. 309). While data generally are available by race and sex, here we restrict our analysis to the age-specific mortality of the total population. (We plan to extend the analysis to population subgroups in the future, but are concerned about extrapolating differentials.) Death rates are available for infants and standard five-year age groups up to age 85, and for age 85 and over. There is reason to be skeptical about measures of mortality at the older ages. With 46% of the population already surviving to

Example 7: Factor Analysis on Mortality Data

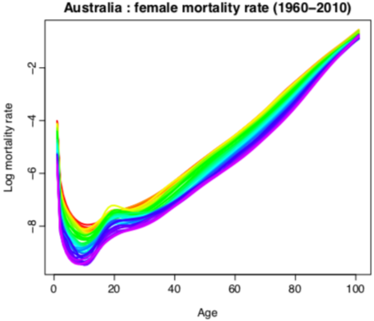
- For the mortality data x_{it} (the death rate for age i at time t), consider the factor model

$$x_{it} = \lambda_{i1} f_{1t} + \lambda_{i2} f_{2t} + \cdots + \lambda_{ir} f_{rt} + \varepsilon_{it}.$$

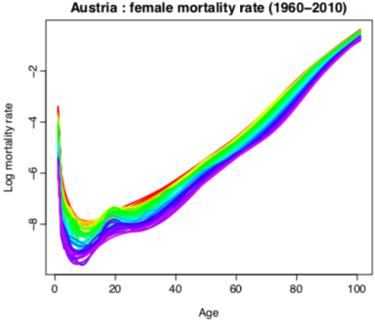
- After obtaining $\widehat{f}_{1t}, \ldots, \widehat{f}_{rt}$, forecasting model is applied on $\widehat{f}_{1t}, \ldots, \widehat{f}_{rt}$ and then get forecasting values $\widehat{f}_{1,t+k}, \ldots, \widehat{f}_{r,t+k}$.
- Forecasting the mortality with

$$\widehat{x}_{i,t+k} = \widehat{\lambda}_{i1} \widehat{f}_{1,t+k} + \widehat{\lambda}_{i2} \widehat{f}_{2,t+k} + \cdots + \widehat{\lambda}_{ir} \widehat{f}_{r,t+k}.$$

Example 7: Multi-Country Mortality Data

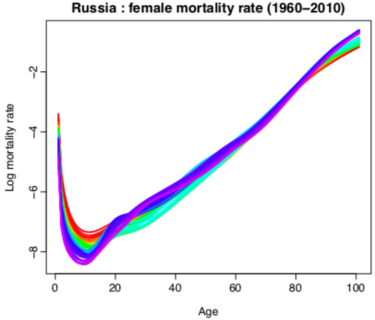


(a) *Female smoothed mortality rates in Australia*

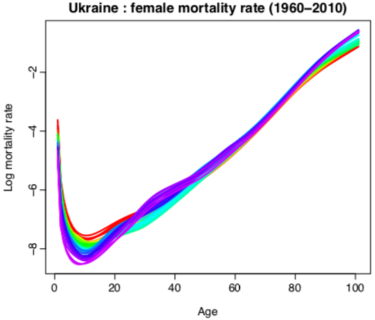


(b) *Female smoothed mortality rates in Austria*

Example 7: Multi-Country Mortality Data



(c) *Female smoothed mortality rates in Russia*



(d) *Female smoothed mortality rates in Ukraine*

Conclusion

- Evaluation of Uncertainty (or Randomness) is the essential problem in Statistical Analysis.
- Determining Optimal or Appropriate Flexibility is the major challenge in Statistical Analysis.
- Complexity of Big Data comes from "curse of dimensionality" and "heterogeneity".
- Advanced Statistical Learning is modern statistics, which provides algorithms and inference for Big Data.