BICKEL AND DOKSUM SUMMARY - VOLUME I

A PREPRINT

Adam Li[1,2]

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, United States
[2]Institute for Computational Medicine, Johns Hopkins University, Baltimore, United States

November 21, 2019

Contents

[1]BD is a hard book to read, so here we try to present a summary of the important concepts in outline format. If you feel like there was an error, please submit an Issue and Pull Request.

List of Figures

List of Tables

1 Useful Notation Reminders

1. Def: A distribution is a parametric distribution if P is in a parametrized class of models: $P \in \mathbb{F} = \{P_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}^d$. Θ is the set of all possibilities of a random variable. d is the dimension of our parameter space.
2. Def: The set of all possibilities of a random variable is Ω
3. Def: The action spaces, $\mathbb{A}$ is the range of the statistical decision procedure. Procedures can include: parameter estimation, hypothesis testing, and confidence region estimation.
4. Def: The empirical distribution is just the
5. Def: The cumulative distribution $F_X(x)$ takes occurrences of the random variable $X = x$ and computes the probability: $P[X \leq x]$.
6. IID: independent and identically distributed according to some probability function (parametric model in our case)

A comment on subscripts

Generally, P is arbitrary except for regularity conditions including, but not limited to:

1. finite second moments: $E_P[X^2] < \infty$ 2. continuity of P

2 Chapter 1: An introduction to important concepts in statistical learning

2.1 Important Concepts and Definitions

Regularity: This means that the stochastic process $\epsilon_n(x) = \sqrt{n}(\hat{F}(x) - F(x)), x \in \mathbb{R}$ converges to a Gaussian process $W^0(F(.))$, which is a Brownian bridge with mean 0 and covariance structure depending on F(.).

2.2 Goodness of Fit and Brownian Bridge:

Problem statement (v1, easy): If we are given a Gaussian distribution, $H : F(.) = \Phi(\frac{-\mu}{\sigma})$ for some μ, σ, then a goodness-of-fit statistic can be:

$$sup_x|\hat{G}(x) - \Phi(x)|$$

$\hat{G}$ is the empirical distribution of $(Z_1, ..., Z_n)$, where each $Z_i = (X_i - \bar{X})/\hat{\sigma}$ is the z-normalized sample point. This G has a null distribution not depending on μ, or σ as a result because it's null is N(0,1). This corresponds to our Z-distribution that we know and love. We compare this to a more general problem.

Problem statement (v2, hard): If we are given a parametric model distribution, $H : X\ P \in \mathbb{F} = \{P_\theta : \theta \in \Theta\}$ is regular, then this problem is very difficult.

2.3 Minimum Distance Estimation

A minimum distance estimate $\theta(\hat{P})$ is the solution to:

$$\theta(P) = argmin\{d(P, P_\theta) : \theta \in \Theta\}$$

where $\hat{P}$, the empirical distribution is substituted for P, and d is some metric defined on the space of probability distributions for X. (i.e. positivity, homogeneity and triangle inequality).

If space X is $\mathbb{R}$, then metrics can act on the Euclidean space. The question of interest is if we can linearized, and generalized to show asymptotic Gaussianness?

2.4 Convergence

There is convergence in the sense of achieving a supremum, or infimum in real analysis. There is also rates of convergence, where the limit happens at a function of a variable.

Def: $\theta(\hat{P})$ convergest to $\theta(P)$ at a rate δ_n if and only if for all $\epsilon > 0$, there exists a $c < \infty$ such that $sup\{P[|\theta(\hat{P}) - \theta(P)| \geq c\delta_n] : P \in M_0\} \leq \epsilon$.

2.5 Permutation Testing

Problem statement: If we are given two samples of data iid: $S_X = \{X_1, ..., X_n\}$ and $S_Y = \{Y_1, ..., Y_m\}$. We can call one the control, and one the treatment from distributions F and G, respectively.

General summary: A permutation test (i.e. randomization test) is a type of statistical significance test, where the distribution of the **test statistic** under the null hypothesis is obtained by calculating all possible empirical values of the test statistic under rearrangements of the labels on observed data points. (i.e. swap X_i, or Y_j into the opposite sets, S_X, or S_Y.)

2.5.1 Fisher's Permutation Test Summary:

$$H_0 : F = G$$
$$H_A : F \neq G$$

We define $g = (g_1, ..., g_n, g_{n+1}, ...g_{n+m})$ is a vector of binary labels assigning each of the observations X_i, Y_j to their original conditions; this changes depending on what we observe obviously.

There are $\binom{(n+m)}{n}$ possible g vectors in general. If H_0 is true, then all these can occur with equal probability. Now, let g^* be the vector of labels that we get from our data sample (S_X, S_Y), $\theta(X)$ be a proposed test statistic, and $\hat{\theta}^* = \hat{\theta}(g^*)$ be the test statistic based on the a specific instance of labeling, g^*.

Our permutation test:

$$P_{perm}[\hat{\theta}^* \geq \hat{\theta}] = \frac{\mathbb{1}\{\hat{\theta}^* \geq \hat{\theta}\}}{\binom{n+m}{n}}$$

Just the number of instances your permuted distribution of test statistics are less than your observed test statistic divided by the total number of possibilities. This is not feasible if the total number of possibilities is large, so instead, we approximate this by choosing **B times** without replacement from the total set of all possible combinations. We then evaluate, and compute $\hat{P}_{perm}$

2.5.2 Choosing B (number of permutations to do):

`https://www.tau.ac.il/~saharon/StatisticsSeminar_files/Permutation%20Tests_final.pdf`

Good notebook: - `https://hasthika.github.io/STT3850/Lecture%20Notes/Ch-3_Notes_students.html`

2.6 Irregular Parameters

TODO

2.7 Stein Estimation

TODO

2.8 Empirical Bayes Estimation

TODO

2.9 Model Selection

3 Chapter 5: Asymptotic Approximations

Analytical forms of the risk function is rare, and computation may involve high dimensional integration.

Either:

1. Approximate risk function $R_n(F) = E_F[l(F, \delta(X_1, .., X_n))]$ with easier to compute and simpler function $\tilde{R}_n(F)$.

2. Use Monte Carlo method to draw independent samples from F using a rng, and an explicit function F. Then approximate the risk function using the empirical risk function. By LLN, if we draw more and more samples, the empirical risk function converges in probability to the true risk function.

3.1 Examples:

3.1.1 Example 1: Risk of the Median

Given $X_1, ..., X_n$ iid F, then we are interested in finding the population median, $\nu(F)$ with estimator: $\hat{\nu} = median(X_1, ..., X_n)$. The risk function for squared error loss is:

$$MSE_F(\hat{\nu}) = \int_{-\infty}^{\infty} (x - F^{-1}(1/2))^2 g_n(x) dx$$

where F is the CDF. $F^{-1}(1/2)$

4 Acknowledgements

AL is supported by NIH T32 EB003383, NSF GRFP, Whitaker Fellowship and the Chateaubriand Fellowship. SVS is supported by NIH R21 NS103113, the Coulter Foundation, Maryland Innovation Initiative, US NSF Career Award 1055560 and the Burroughs Wellcome Fund CASI Award 1007274. Computational resources were also provided by the Maryland Advanced Research Computing Center (MARCC).

5 Supplementary Material

References