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 24, 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

1. 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. Bias of an estimator: This is the difference in expectation of your estimator to the true parameter value in the population model.
3. Variance: This is the variance in your estimator.
4. MLE: maximum likelihood estimator is an estimator that maximizes a likelihood function
5. Estimator: A function that takes a sample of data (i.e. instance of a random variable) and produces a value in Θ, your parameter space of interest.
6. Loss function: A function that takes your estimated parameter, $\hat{\theta}$ and true parameter, θ, and produces a number in $\mathbb{R}_+$ (i.e. loss is non-negative).
7. Risk functional: A functional that takes your loss function, and produces a "risk" of your estimator. The risk would be the expectated value of your loss function under the true model. $E_P[l(\hat{\theta}, \theta)]$. If loss is squared error, then risk functional is commonly known as mean-squared error.
8. admissability: There are infinitely many possible estimators for any problem. However, you want to have a principled way of choosing an estimator if you have multiple proposed estimators. Let us say f, and g are proposed estimators for true value θ. Then an estimator g is inadmissable if $E_\theta[l(g, \theta)] \leq E_\theta[l(f, \theta)]\ \forall \theta \in \Theta$. That is, for every possible value of the parameter, if g's risk is greater then another estimator, then you should never use g as an estimator for θ.

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

2.6 Irregular Parameters

TODO:
1. An explanation of the regular model vs irregular model issue.
2. Explain why histogram estimate versus parameter estimation in the parametric model setting.

Problem illustration
Consider histogram estimate of a one-dimensional density p(.). That is:

$$\hat{p}(t) = \hat{P}[\mathbb{1}_j(t)]/h$$

where $\mathbb{1}_j = (jh, (j+1)h)$, is the interval that contains data samples t. It is also the unique interval, and h is the size of the interval. This is a **plug-in estimate** for the parameter $p_h(t) = P[\mathbb{1}_j(t)]/h$, which is the true population density for some bin sizes h. Note that the only change occurred in the probability model $\hat{P}$ to P. One is the estimate, one is the truth.

$\hat{p} \neq p$ for h > 0, but if we take $h = lim_{n->\infty} h_n = 0$, then $\hat{p} \to p \; \forall t$. That is, as the size of the intervals (i.e. bins) approaches 0, then the plug-in estimate approaches the true density. Essentially, we get a few properties asymptotically for the "bias" and "variance" of the plug-in estimator:

$$E_P[\hat{p}_h(t) - p_h(t)] = 0$$

$$Var_P[\hat{p}_h(t)] = (p_h(t) - p_h(t)hp_h(t))/hn$$

The bias of the h parameter (i.e bin size) is given by:

$$Bias(h) = \frac{1}{h}\int_{jh}^{(j+1)h}(p(s) - p(t))ds$$

is the integral form of the expectation, and goes to 0 when h -> 0. On the other hand, the variance by limit analysis of h and n, only goes to 0 if h goes to 0 slower then n goes to ∞. So there is a balance here between having h go to 0 as fast as possible (lowers the bias), versus having it go slower then the rate of n (lowers variance). This is essentially a view of the bias-variance tradeoff that is common in statistical methods.

2.7 Stein Estimation

Here, BD considers a very specific example of the analysis of variances in a p-sample Gaussian model.

$X = \{X_{ij} : 1 \leq i \leq n, 1 \leq j \leq p\}$, with X_{ij} independent samples distributed from $N(\mu_j, \sigma_0^2)$ parametric model. Note that j is the index for which normal distribution mean we use, and i is the sample index. σ_0^2, the population variances are assumed equal and *known* with $\mu_p = (\mu_1, ..., \mu_p)$ unknown. Then X $P(n, p)$ for this class of distributions. The MLE of μ_p is:

$$\bar{X}_p = (X_{\cdot 1}, ..., X_{\cdot p})$$

which is the sample mean for each group of samples. $X_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$.

2.8 Empirical Bayes Estimation

TODO

2.9 Model Selection

TODO

3 Chapter 2: Methods of Estimation

In general, there is a random variable of interest X $P \in \mathbb{P} = \{P_\theta : \theta \in \Theta\}$. Now, we want to estimate θ in some reasonable manner with functions $\hat{\theta}$ based on the vector of observations, X. Our goal is to make this estimator somehow close to the true θ.

3.1 Heuristics in Estimations

Definition 3.1. Contrast Function $\rho : X \times \Theta \to \mathbb{R}$ is a function that takes the random variable distribution and the parameter space to a real number. This is known as the contrast function.

and a discrepancy function based on the population is defined as:

Definition 3.2. Population Discrepancy $D(\theta_0, \theta) = E_{\theta_0}[\rho(X, \theta)]$ is the expected value of the contrast based on the true value θ_0. θ_0 is the unique minimizer of D.

If, P_{θ_0} were the true model, and we knew $D(\theta_0, \theta)$, then we could obtain θ_0 as the minimizer. However, since we do not know D, we instead try to minimize $\rho(X, \theta)$, which would be $\hat{\theta}(X)$ to estimate θ_0.

Definition 3.3. Minimum Contrast Estimate $p(.,.)$ is a contrast function and $\hat{\theta}(X)$ is a minimum contrast estimate of the true θ_0.

Euclidean Space
When we are operating in finite Euclidean space, the true θ_0 is an interior point of our parameter space and the discrepancy function is smooth, then we would expect that the gradient of our discrepancy is equal to 0 when evaluated at the minimum, $\theta = \theta_0$.

$$\nabla_\theta D(\theta_0, \theta)|_{\theta=\theta_0} = 0$$

So, as we did earlier, we do not know D, so we use a plug-in of it with $\rho(X, \theta)$ instead. So we are interested in solving equations of the form:

$$\nabla_\theta \rho(X, \theta) = 0$$

which is known as a form of ***estimating equation***.

In more generality than Euclidean space
Now, say we are given a general function of the form:

$$\Psi : X \times \mathbb{R}^d \to \mathbb{R}^d$$

and

$$V(\theta_0, \theta) = E_{\theta_0} \Psi(X, \theta)$$

If V = 0 has $\theta = \theta_0$ as its unique solution for all $\theta_0 \in \Theta$, then we say $\hat{\theta}$ solving the equation $\Psi(X, \hat{\theta}) = 0$ is an estimating equation estimate. Note here that θ is our variable, and θ_0 is some fixed parameter that is the "truth".

3.1.1 Examples

Minimum contrast estimates are very abstract at first glance, so it is worthwhile to look at some examples you may already be familiar with to put in the context of minimum contrast estimators.

Least Squares Estimation

If we have $\mu(z) = g(\beta, z), \beta \in \mathbb{R}^d$, with g known. The data $X = \{(z_i, Y_i) : 1 \leq i \leq n\}$, where $Y_1, ..., Y_n$ are independent labels. There are many choices for how we can frame this as minimum contrast, but here we define the following:

The **contrast function** $\rho(X, \beta)$ is squared Euclidean distance between vector Y and the vector expectation of Y, $\mu(z) = (g(\beta, z_1), ..., g(\beta, z_n))$. Specifically it looks like:

$$\rho(X, \beta) = |Y - \mu|^2 = \sum_{i=1}^{n} [Y_i - g(\beta, z_i)]^2$$

The discrepancy function is:

$$D(\beta_0, \beta) = E_{\beta_0} \rho(X, \beta) = n\sigma_0^2 + \sum_{i=1}^{n} [g(\beta_0, z_i) - g(\beta, z_i)]^2$$

this is minimized when $\beta = \beta_0$ and is unique minmizer **if and only if** the parametrization is identifiable.

The contrast function is minimized here:

An estimate $\hat{\beta}$ that minimizes $\rho(X, \beta$ exists if $g(\beta, z)$ is continuous in β and that $lim\{|g(\beta, z)| : |\beta| \to \infty\} = \infty$.

With differentiable functions If $g(\beta, z)$ is differentiable in β, then $\hat{\beta}$ satisfies: $\nabla_\theta \rho(X, \theta) = 0$. Then it makes the system of estimating equations:

$$\sum_{i=1}^{n} \frac{\partial g}{\partial \beta_j}(\hat{\beta}, z_i) Y_i = \sum_{i=1}^{n} \frac{\partial g}{\partial \beta_j}(\hat{\beta}, z_i) g(\hat{\beta, z_i})$$

with $1 \leq j \leq d$. If g is linear, it is a summation of z_{ij} multiplied by their slopes, β_j. These can be used to derive the normal equations, which can be wrriten in matrix form to solve least-squares. Least squares in this sense, are just a specific instance of minimum contrast.

Method of Moments (MOM) TODO

3.2 Plug-in and Extension Principle

In the case of iid situation, there are two principled heuristics that can be used to estimate parameters. One is the plug-in principle, and the next one is the extension principle.

3.2.1 Plug-in Principle

This is just an abstract way of saying we plug in the empirical distribution we see into our estimator, as a "plug-in" estimate for our parameter. This is justified via the law of large numbers.

3.2.2 Extension Principle

This is just an abstract way of saying that when we have an estimator, ν on a submodel of our proposed probability model, then a new estimator, $\bar{\nu}$ on the full model of our proposed proability model is an extension of ν. It must have the property that $\bar{\nu}(P) = \nu(P)$ on the submodel.

3.2.3 Examples of Plug-in and Extensions

Example - Frequency Plug-In and Extension

Hardy Weinberge TODO

3.3 Minimum Contrast Estimates

3.4 Maximum Likelihood in Exponential Families

4 Chapter 3: Measure of Performance and "Notions" of Optimality in Estimation Procedures

In general, there is a random variable of interest X $P \in \mathbb{P} = \{P_\theta : \theta \in \Theta\}$. Now, we want to estimate θ in some reasonable manner with functions $\hat{\theta}$ based on the vector of observations, X. Our goal is to make this estimator somehow close to the true θ.

5 Chapter 4: Hypothesis Testing and Confidence Regions

6 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.

6.1 Examples:

6.1.1 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. $F^{-1}(1/2)$

7 Acknowledgements

AL is supported by NIH T32 EB003383, NSF GRFP, Whitaker Fellowship and the Chateaubriand Fellowship.

8 Supplementary Material

References