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

December 6, 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 - Edition 1

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. $X_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$.

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

3.1 Important Concepts and Definitions

1. "Regular" Models
2. Sufficiency
3. Minimal Sufficiency
4. Admissability / Inadmissability
5. Parametric models
6. Order statistics
7. Empirical distribution function
8. Glivinko Cantelli Bound
9. Hoeffding Bound
10. Gauss-Markov Theorem

Definition 3.1. A Statistic $T : X \times \mathbb{T}$ is a function that takes the sample space and maps to some possible values of statistics, usually Eucliean $\mathbb{R}^d$ space, where d is the dimensionality of the statistic.

Well-known examples of statistics are the sample mean, and sample variance, but they can be any arbitrary mapping from sampled data.

Definition 3.2. The empirical distribution function $\hat{F}(X_1, ..., X_n; x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)$ which basically tells us the probability that our sampled data is less then certain discrete values x. This essentially "bins" our data based on the indicator function.

This statistic is nice beause it is easy to compute, and also we know asymptotically approaches the true F, as we take $n \to \infty$.

Definition 3.3. Regular models For any parametric model, it is considered a "regular parametric model", as long as either:

1. Continuous: All P_θ are continuous with corresponding densities $p(x, \theta)$.
2. Discrete: All P_θ are discrete with frequency functions $p(x, \theta)$ and there exists a countable set $\{x_1, x_2, ...\}$ that is independent of θ such that the normalizing property is achieved for all θ (i.e. $\sum_{i=1}^{\infty} p(x_i, \theta) = 1$)

Pretty much this just is BD way of saying from now on, regular parametric models are either densities or frequency functions, but these are just the "joint probabilities" that we are used to seeing.

3.2 Decision Theory Framework

The premise of this section is to define a rigorous framework to think about how to make **decisions using data** in an optimal sense. In the real world, we pretty much never have access to the true population parameters, and so we have to make **assumptions on the model that fits the population**, and generally we use parametric models. Then, the goal becomes fitting these parametric models with data and choosing the best possible estimators we can derive as functions of data. In this aspect, we then must define various objects:

Definition 3.4. The action space A is an action space that consists of all actions, or decisions, or claims that we can make given a new "data, or component".

Examples:

1. the real number line, denoting mean of male heights
2. the set of 0,1, denoting if we see disease state or not
3. estimation, hypothesis testing, ranking and prediction are all results of an "action space"

Definition 3.5. The loss function $l : P \times A \to \mathbb{R}_+$ is a function that takes the true model, and compares the output action (e.g. estimate) and produces a non-negative real number.

Examples:

1. quadratic loss (i.e. l2 loss)
2. l1 loss
3. cross-entropy loss
4. 0-1 loss
5. 0-a-b loss

Definition 3.6. The decision rule $\delta : X \to A$ is a function that acts on our samples to produce an action. (e.g. a sample estimate of a parameter). This is just a "generalization" of an "estimator" because it covers everything from estimation, hypothesis testing, ranking and prediction.

Definition 3.7. The risk function $R : P \times X \to \mathbb{R}_+$ is the risk function that determins the expected value of our loss over the entire sample space, for a specific true model, P.

$R(P, \delta(.)) = E_P[l(P, \delta(X))]$, which measures the performance of the decision rule. Note why this is important. Loss of your decision rule is only for your specific samples, but risk is the expected loss over entire sample space, which is what we actually care about (think training vs testing vs validation data).

Now, the risk function can generally be very complicated, but if we consider l2-loss, then our risk function is the well-known **mean-squared error** (i.e. MSE). This then allows the decomposition of risk into the well-known **Bias and Variance**! Consider, $\hat{\theta}$ as a decision rule estimator for a true parameter, θ, which parametrizes a parametric model P.

$$MSE(\hat{\theta}) = R(P, \hat{\theta}) = E_P[(\hat{\theta}(X) - \theta(P))^2] \tag{1}$$
$$= E_P[(\hat{\theta}(X) - E[\hat{\theta}] + E[\hat{\theta}] - \theta(P))^2] \tag{2}$$
$$= Bias(\hat{\theta})^2 + Var[\hat{\theta}] \tag{3}$$

The nice thing about MSE is that it's generally computable "easily", and it has some nice connections when we use Bayesian statistics. But generally if you think about it, optimizing for the average performance of an estimator might not be what you want. Consider in finance, perhaps you want to minimize the worst case scenario, then the loss would actually be the $l - \infty$ loss potentially, rather then l2 loss.

3.3 Ways of Comparing Decision Procedures - Based on risk

Now, that BD has defined the necessary components of a rigorous decision theory framework, one might ask: How can we compare possible estimates in a principled way? Part of the art in statistics is choosing the best "metric of comparison" for your problem. MSE is not the best risk functional for all problems, although it is a nice one to start with potentially.

1. Inadmissable versus admissable: If we can determine that a decision rule has better risk for all possible parameter values, then we would surely use this one. This is in general hard to verify though. Note that Wald shows that all admissible procedures are Bayes procedures! (so checking Bayes is sufficient for checking admissability)
2. Bayes optimality: Here, we are interested in obtaining decision procedures that improve upon a risk only for some subset of our parameter space that is governed by a prior.
3. Minmax optimality: Here, we optimize decision procedures based on the worst possible risk they could have.
4. Unbiased optimality: Here, we restrict our decision procedures to have unbiased property (i.e. expectated value is the true parameter), but note that there can be incredibly high variance as seen in the bias/variance decomposition of MSE
5. Randomized procedures: **not sure, how to explain this**

3.4 Sufficient Statistics, Rao-Blackwellization, and Neyman-Pearson Factorization

In order to understand some of these theorems and concepts it would be useful to remind yourself of the following theorems/concepts:

1. Holder's inequality for Normed Linear Spaces, Inner product spaces, and Measureable spaces

2. convexity of a set and convexity of functions
3. continuity of a function for open versus closed sets
4. Cauchy sequences and their relation to compactness and their relation to continuity and boundedness

Definition 3.8. Sufficiency T(X), $T : X \to \mathbb{T}$ is a sufficient statistic if the conditional distribution of sample space, X given T(X) = t is independent of parameter, θ. In other words: $p(X|T(X) = t) \neq f(\theta)$, where θ parametrizes our parametric model P.

Definition 3.9. Minimal Sufficiency T(X), $T : X \to \mathbb{T}$ is a minimal sufficient statistic if.

Theorem 1. Neyman-Pearson Factorization Theorem This is a way of proving that a statistic is sufficient because it is a necessary and sufficient condition in regular models.

Theorem 2. Rao-Blackwell Theorem This is a way of improving the MSE risk of a model given that you have a sufficient statistic. Note this does not guarantee you improve. Note that this is also only a result for MSE, not any other risk functional. However, it can be generalized to convex loss functionals.

3.5 Example Problems and Solutions - Chapter One

3.5.1 Bayes estimator for Bernoulli Trials

3.5.2 Minimal Sufficiency Derived from Neyman-Pearson Factorization Theorem

3.5.3 The Order Statistics are Sufficient

3.5.4 The Order Statistics are Equivalent to the Empirical CDF

3.5.5 The Minimal Sufficient Statistic for a Laplace Model

3.5.6 Explanation: Suffficiency is important in the Rao-Blackwell Theorem

4 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.

4.1 Heuristics in Estimations

Definition 4.1. 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 4.2. 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 4.3. 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".

4.1.1 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) MOM estimates are equations that **estimate** say, d moments of the population with their corresponding sample estimates. So we **assume the existence** of the following d population moments:

$$\mu_j(\theta) = \mu_j = E_\theta[X^j], 1 \leq j \leq d$$

The sample moments are:

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j, 1 \leq j \leq d$$

The next step is to express our parameter of interest, θ, as a continuous function of the first d moments. So if we assume that the mapping of $\theta \to (\mu_1(\theta), ..., \mu_d(\theta))$ is 1-1, then we can estimate θ by the d equations:

$$\hat{\mu}_j = \mu_j(\hat{\theta}), 1 \leq j \leq d$$

That is, we express the parameters as a function of the population moments, and then use the plug-in principle by **plugging in** the sample moments instead. This then gives us the MOM estimate of the parameters. Note that the MOM estimate is not unique.

Solving MOM Equations
In the MOM framework, someone usually sets up d moment equations, with possibly nonlinearities. Then to find a solution is equivalent to finding a root of the optimization equation, which we can use any number of zero-finding algorithms including but not limited to: Line Search, Trust-Region, or Newton-Raphson algorithms.

4.2 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.

4.2.1 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.

4.2.2 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.

4.2.3 Examples of Plug-in and Extensions

Example - Frequency Plug-In and Extension

Hardy Weinberg TODO

4.3 Minimum Contrast Estimates

Maximum likelihood only **makes sense in regular parametric models**. The likelihood function of θ is $L_x(\theta)$, which is just the density as a function of θ for fixed x, samples. The method of MLE is to find an estimate $\hat{\theta}(x)$ that is "most likely" under the likelihood function to have produced the data x.

$$L_x(\hat{\theta}(x)) = p(x, \hat{\theta}(x)) = max\{L_x(\theta) : \theta \in \Theta\}$$

In the context of Bayes estimation, the MLE is the mode of the posterior distribution if Θ is finite and the prior is a uniform distribution.

4.3.1 MLE as Minimum Contrast Estimates from the Kullback-Leibler Divergence

If we consider the likelihood function: $L_x(\theta)$, and we take its logarithm: $l_x(\theta) = logL_x(\theta)$, then if the MLE exists, then it will minimize $-l_x(\theta)$ because the negative log is a strictly decreasing function. Now, if we **assume** that the samples are independent with some density function $f(x, \theta)$, then the log-likelihood function looks like:

$$l_X(\theta) = log \prod_{i=1}^{n} f(X_i, \theta) = \sum_{i=1}^{n} log(f(X_i, \theta))$$

This now, is a random variable that is the sum of independent random variables.

MLE as an Information Quantity
Note that, if we define a contrast function, $\rho(x, \theta) = -l_x(\theta)$ and $\hat{\theta}$ as a minimum contrast estimate, then our discrepancy function is:

$$D(\theta_0, \theta) = -E_{\theta_0}[log(p(X, \theta))]$$

where D is uniquely minimized when $P_\theta = P_{\theta_0}$, which is equivalent to:

$$D(\theta_0, \theta) - D(\theta_0, \theta_0) = -(E_{\theta_0}[log(p(X, \theta))] - E_{\theta_0}[log(p(X, \theta_0))]) \tag{4}$$

$$= -E_{\theta_0}[log\frac{p(X, \theta)}{p(X, \theta)}] > 0 \tag{5}$$

$$\tag{6}$$

In information theory, $D(\theta_0, \theta_0)$ is the entropy of X, which can also be seen as the smallest value that D can take. Thus, $D(\theta_0, \theta) - D(\theta_0, \theta_0)$ is equivalent to the Kullback-Leibler (KL) information divergence between the densities p_θ, p_{θ_0}. We can replace the expectation with a summation when X is discrete, or integrals when X is continuous. From information theory, we also know that KL divergence is always ≥ 0 and is equal to 0, only when $P_1(x) = P_0(x)$.

Summary:
Since MLE was shown to be a minimum contrast estimate, and the contrast function defined above satisfies the condition of being a contrast function, then we can use the extension principle, and show that the MLE is in fact minimizing the KL divergence between the empirical probability $\hat{P}$ and the model probability P_θ.

4.3.2 MLE as Estimating Equations (i.e. Method of Moments)

If Θ is open (a solution cannot be on the boundary) and $l_X(\theta)$ is diffierentiable, and $\hat{\theta}$ exists, then the estimating equation:

$$\nabla_\theta l_X(\theta) = 0$$

solves for $\hat{\theta}$.

4.4 Maximum Likelihood in Exponential Families

We know that MLE does not always exist, nor is it unique. However, in a very special class of models, they always exist and are unique. The canonical exponential families have unique maximum likelihood estimates. This section's main theorem is Theorem 2.3.1, which shows the existence of uniqueness of the MLE. I restate it here for brevity.

Theorem 3 (Existence and Uniquenee of MLE in Canonical Exponential Families and Relation to Log-Partition Function)**.** Suppose $\mathbb{F}$ is the canonical exponential family, with T as the sufficient statistic, h as the base measure, $\mathbb{E}$ as the natural parameter space, and ν as the natural parameter. If the following conditions hold:

1. $\mathbb{E}$ is open
2. $\mathbb{F}$ is rank k
3. If $t_0 = T(x) \in \mathbb{R}^k$ satsifies the equation: $P[c^T T(X) > c^T t_0] > 0, \forall c \neq 0$

then the MLE, $\hat{\nu}$ exists, and is unique, and it is a solution to the equation:

$\dot{A}(\nu) = E_\nu[T(X)] = t_0$

If $t_0 = T(x)$ does not satisfy the equation: $P[c^T T(X) > c^T t_0] > 0$, then the MLE does not exist, and $\dot{A}(\nu) = E_\nu[T(X)] = t_0$ has no solution.

Here, I comment on various properties of the theorem and why they are important:

1. $\mathbb{E}$ must be open in order for the parameter set to not include boundary points. We saw that at the boundary, irregularities can occur when performing maximum likelihood estimation.
2. the exponential family model needs to be of rank k, because that is the dimension of the sufficient statistic. If it was of lesser rank, then T would not be a sufficient statistic
3. This last condition is necessary because when the math is worked out, this condition is necessary to prove the theorem

4.5 Making sense of Plug-in estimates, Minimum Contrast estimates and Maximum Likelihood estimate

Here, when we reviewed plug-in principles using the simple empirical distribution statistics, it is important to remember that it is simply a **heuristical** procedure, which are i) easy to compute and ii) generally lead to good first estimation procedures.

The same is true for method of moments, where we can write down the estimating equations for "d" moments of the population. Note, that here we also do a "plug-in", where we substitute instead of the population moment, the sample moments, in order to actually be able to estimate anything.

When we consider the general framework of minimum contrast estimates, then we see that many "well-known" things fall under this framework: least-squares estimators (minimizing the contrast function mean-squared error), maximum likelihood estimators (minimizing the contrast function related to entropy and KL divergence).

MLE seems like a good first approach and it IS for canonical exponential families because Thm. 3 shows that MLEs exist and are unique for canonical exponential families with certain regularity conditions. In **addition**, the theorem shows an important relationship with the log-partition function A(.). Remember that in the general sense, the A(.) is a normalizing constant that make the model a probability density function; in Bayesian statistics, when you do Markov Chain Monte Carlo sampling, or variational inference to obtain an approximation to the posterior distribution, it is always because it is "hard" to determine the normalizing constant. Here in canonical EF though, we have that the log-partition function gives the first two moments of the model, i.e. the expected value and the variance.

4.6 Example Problems and Solutions - Chapter Two

4.6.1 2.3.7

4.6.2 MLE as a generalized MoM Estimator

4.6.3 Comparison of MLE and MoM Estimators on Finite-sample Gamma for MSE as our Risk Functional

5 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. There are different things we can consider from, ease of computation, consistency, robustness, or minimizing expected risk, etc.

In Chapter 2, we saw that plug-in estimates and method of moments are easy to compute. In Chapter 5, we will see that the MLE for canonical EF are consistent and we will also define consistency then.

5.1 Bayes Optimality

Problem setup:
We are given a parametric model family $\mathbb{F} = \{P_\theta : \theta \in \Theta\}$, with an action space, A, and a loss function $l(\theta, a)$.

Then if we sample idd data X P_θ, and specify a decision procedure δ that can either be randomized, or not, then we define the risk function: $R(., \delta) : \Theta \to \mathbb{R}^+$

$$R(\theta, \delta) = E_\theta[l(\theta, \delta(X))]$$

The risk is a measure of performance of your decision rule δ for this SPECIFIC model. At the very least, you do not want to consider inadmissable estimators (i.e. risk is worse for every value of θ). In the Bayes context, we introduce a prior density π for the parameter θ. Then we can consider the following **Bayes risk**.

$$r(\pi, \delta) = E[R(\theta, \delta)] = E[l(\theta, \delta(X))]$$

Definition 5.1. Minimum Bayes Risk $R(\pi) = inf\{r(\pi, \delta) : \delta \in D\}$ is the minimum Bayes risk of the problem.

Our goal is to identify Bayes rules, δ^*_π, such that: $r(\pi, \delta^*_\pi) = R(\pi)$.

5.1.1 Selection of Priors π

Improper priors

Uniform/constant priors

Jeffrey's priors

Conjugate priors

5.1.2 Bayes Estimation for Squared Error Loss

We are interested in estimating $q(\theta)$.

In our setup, we now constrain our loss to be the quadratic loss function: $l(\theta, a) = (q(\theta) - a)^2$ using a nonrandomized decision rule, δ. Consider $\pi(\theta)$ as our prior on the random vector θ. We want now to find the function $\delta(X)$ that minimizes $r(\pi, \delta) = E[q(\theta) - \delta(X)]^2$.

This boils down to either the Bayes risk being ∞ for all δ, or that we arrive at the Bayes rule, $\delta^*(X) = E[q(\theta)|X]$, which is just the **mean of the posterior distribution!**

5.1.3 Bayes Estimation for General Loss Functions

We are interested in estimating $q(\theta)$.

In our setup, we now consider general loss functions $l(\theta, a)$ using a nonrandomized decision rule, δ. Consider $\pi(\theta)$ as our prior on the random vector θ. We apply the same idea to formulate the posterior risk, which is:

$$r(a|x) = E[l(\theta, a)|X = x]$$

5.2 Minimax Optimality

Minimax optimality is considering the "maximum", or supremum of possible risks over the space of parameter values. It is used in optimizing for the worst-case outcome, but in many cases are shown to be inadmissable!

5.3 Unbiased Optimality

5.3.1 Fisher Information (Matrix, or Value)

5.3.2 The Information Inequality Provides a Lower Bound on the Variance of Your Sufficient Statistic

5.4 Computation and Interpretability

5.5 Robustness

5.5.1 Gross Error Models

5.5.2 Sensitivity Curves

6 Chapter 4: Hypothesis Testing and Confidence Regions

7 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.

7.1 Examples:

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

8 Inference in Multiparameters

From chapters 2-5, we explored the behavior of estimates, tests and confidence regions just for simple regular one-dimensional parametric models. Now, we want to extend that understanding to d-dimensional models.

8.1 Inference for Gaussian Linear Models

There are n Y_i independent observations has a distribution that depends on known constants $z_{i1}, ..., z_{ip}$. The setup of any Gaussian linear model is in summation form and matrix form:

$$Y_i = \sum_{j=1}^{p} z_{ij}\beta_j + \epsilon_j = z_i^T\beta + \epsilon_i,\ i = 1, ..., n$$

with ϵ_i iid noise samples from $N(0, \sigma^2)$. Y_i is called the response variable, and z_{ij} are the design values, which form the design matrix. Given samples Y, and a matrix form of Z, and assuming that noise is distributed as such, then this is traditionally solved using least-squares (which can be proven to be optimal).

8.1.1 One-Sample Location

If there is just a single mean:

$$Y_i = \beta_1 + \epsilon_i$$

then the solution $\beta_1 = E[Y]$.

8.2 Canonical Form of the Gaussian Linear Model

8.3 Estimation for Gaussian Linear Models Parameters

8.4 References

https://www.stat.berkeley.edu/ aditya/resources/LectureSIX.pdf https://www.stat.berkeley.edu/ aditya/resources/LectureSEVEN.pdf http://www.stat.cmu.edu/ larry/=stat401/lecture-21.pdf

9 Acknowledgements

We thank Carey Priebe for a great course in understanding the **concepts** of basic parametric statistical learning theory and how it is related to everyday research.

10 Supplementary Material

Bickel and Doksum book is in the repository for this summary.

The DGL book will be next on our list to summarize.

References