EE290 Mathematics of Data Science Lecture 16 & 17 - 10/24/2019

Lecture 16 and 17: Adaptive Estimation

Lecturer: Jiantao Jiao *Scribe: Kuan-Yun Lee and Baihong Jin*

What does it mean to do adaptive estimation, and why should we care about it? There are many ways to answer this question—in this lecture, we will look at this from a minimax risk/regret perspective. The scribe is structured in the following way. First, we will give a short review on the definitions of statistical experiments and risk. Then, we will motivate an example where it is beneficial for us to have adaptive estimators. Finally, we will introduce the James-Stein (JS) estimator and show that it adaptively achieves the minimax risk in Pinsker's theorem.

1 Statistical Experiments and Minimax Risk

In order to set the stage for our lecture, we begin by introducing general terms for *statistical experiments* and minimax risk. We will follow notations used in class and also borrow notations from the statistical minimax literature (see, e.g., [5]).

Consider a family of probability distributions $\mathcal{P} = \{P_\theta\,;\,\theta \in \Theta\}$ with support $\mathcal{X}$. We consider θ as a mapping from $\mathcal{P} \to \Theta$, such that for any $\mathbb{P} \in \mathcal{P}$ there is a corresponding parameter $\theta(\mathbb{P})$. Given an unknown fixed distribution $\mathbb{P} \in \mathcal{P}$, we observe a random variable $X \sim \mathbb{P}$. Based on this observation we choose an action $a(X)$ to estimate $\theta^* := \theta(\mathbb{P})$. For a realization $X = x$, We associate a loss function $l(a(x), \theta)$ that characterizes the distance between our estimation $a(x)$ and the true parameter θ^*. This loss function help us assess how well we can perform an estimation on a single parameter θ^* and a single realization of $X = x$. In order to see how well our action function $a(x)$ performs for the parameter θ, we consider the expectation over $l(a(X), \theta)$,

$$\text{Risk}\,(a(\cdot), \theta)) \triangleq \mathbb{E}\; l(a(X), \theta). \tag{1}$$

where the expectation is taken over the samples X. This is known as the *risk function* associated with parameter $\theta \in \Theta$ and action function $a : \mathcal{X}^n \to \Theta$.

An interesting question to ask is how one can evaluate the performance of an action function $a(x)$ over the parameter space Θ. A common evaluation method is to find the worst-case risk $\sup_{\theta\in\Theta} \mathbb{E}\; l(a(X), \theta)$. Under this worst case risk setting, one can then find the action function that minimizes this worst case risk by solving the optimization function $\inf_{a(\cdot)} \sup_{\theta\in\Theta} \mathbb{E}\; l(a(X), \theta)$. This function is called the *minimax risk* in the literature, and in this lecture we will use the notation

$$\text{Minimax Risk} \triangleq \inf_{a(\cdot)} \sup_{\theta\in\Theta} \underbrace{\mathbb{E}\, l(a(X), \theta)}_{\text{Risk}(a(\cdot),\theta))} \triangleq R^*(\Theta). \tag{2}$$

2 Motivating Adaptive Estimation

Now, let's motivate adaptive estimation by looking at the following problem. Consider a sequence of parameter spaces $\mathbb{S} := \{\Theta_1, \Theta_2, \ldots\}$ that satisfy

$$\Theta_1 \subset \Theta_2 \subset \Theta_3 \ldots. \tag{3}$$

Is it possible to find a *single* action function $a(x)$, such that regardless what the true parameter space $\Theta^* \in \mathbb{S}$ is, its worst case risk over Θ^* is always close to the minimax risk of the true parameter space Θ^*? In other words, can we find action $a(x)$ such that for all $i = 1, 2, \ldots$,

$$\sup_{\theta\in\Theta_i} \; \mathbb{E}\; l(a(X), \theta) \approx R^*(\Theta_i)? \tag{4}$$

Why is such an action function potentially valuable for us?

Let's consider the following simple example. Define the sets

$$\Theta_c \triangleq \left\{ \theta : \sum_{i=1}^{n} \theta_i^2 \le c^2 \right\} \tag{5}$$

and assume we observe Y based on the Gaussian noise model,

$$Y = \theta + \sigma_n \cdot \mathcal{N}(0, I_n) = \mathcal{N}(\theta, \sigma_n^2 I_n). \tag{6}$$

The true parameter space Θ_c^* is assumed to be unknown, and we would like to estimate the parameter vector θ based on the observations Y and the loss function $l(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2 = \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2$.

The first natural estimator which we call the *means estimator* is to guess

$$\hat{\theta} = Y. \tag{7}$$

In this case, for any parameter vector θ we can calculate the risk as

$$\mathbb{E}_\theta \; l(\hat{\theta}, \theta) = \mathbb{E}\|Y - \theta\|_2^2 = \sigma_n^2 \cdot n. \tag{8}$$

This implies that this estimator has the same hardness for any true parameter space Θ^*. But this can't be optimal—for example, if the true parameter space is Θ_r where r is very close to 0, we would expect the risk of an optimal estimator to be close to 0 as well, since we could just always use a naïve estimator 0 and still be reasonable.

3 Oracle Inequality (Competitive Inference)

There is a more general view of adaptive estimation in the language of competitive inference. We assume that there exists an oracle that knows more about the underlying true distribution than us, and our goal is to design estimators whose performance is as close as possible to that of the oracle for all problem instances. We now show through examples that choose different oracles leads to drastically different optimality criteria, and it is usually up to the statistician to select the most appropriate oracle in practical applications.

Consider a discrete distribution $p = (p_1, p_2, \ldots, p_k)$ and we obtain n samples $X_1, X_2, \ldots, X_n \sim p$. The maximum likelihood estimator for p is the empirical frequencies:

$$\hat{p}_i \triangleq \frac{1}{n} \sum_{i=j}^{n} \mathbb{I}(X_j = i). \tag{9}$$

If we use the ℓ_1 loss to evaluate the estimation error, the risk associated with $\hat{p}$,

$$\mathbb{E}\ell_1(p, \hat{p}) = \sum_i \mathbb{E}\,|p_i - \hat{p}_i| \tag{10a}$$

$$= \sum_i \sqrt{\mathbb{E}\,|p_i - \hat{p}_i|^2} \tag{10b}$$

$$\le \sum_i \sqrt{\frac{p_i(1-p_i)}{n}} \tag{10c}$$

$$\le \sum_i \sqrt{\frac{p_i}{n}} \tag{10d}$$

$$\le \sqrt{\frac{k}{n}} \tag{10e}$$

It turns out that the rate $\Theta(\sqrt{k/n})$ is tight [1]. We can view this minimax risk result as the performance gap between our estimator and the oracle that knows *exactly* the true distribution p, which is a too strong oracle: indeed, we need number of samples $n \gg k$ to make the performance gap vanish.

It turns out that if we find a less powerful oracle, then we not only can compete with it (the performance gap will vanish as long as $n \gg 1$), but also demonstrate the empirical estimator is no longer optimal in this new formulation.

Consider an oracle that has access to the *sorted* version of p, but it does not know how to assign these probabilities to each symbol. Equivalently, we can assume the oracle has access to p, but the samples it sees come from a permuted version of p, denoted as $\pi(p)$, and the oracle is required to estimate $\pi(p)$ accurately under ℓ_1 loss.

Mathematically, the performance gap between our estimator $\hat{p}_s$ and the oracle is defined as

$$\inf_{\hat{p}_s} \sup_{p \in \mathcal{M}_k, k \geq 1} \left(\mathbb{E}\ell_1(\hat{p}_{\text{s}}, p) - \inf_{\hat{p}_{\text{oracle}}} \sup_{\pi} \mathbb{E}\ell_1(\hat{p}_{\text{oracle}}, \pi(p)) \right), \tag{11}$$

where $\mathcal{M}_k$ denotes the space of discrete distributions with alphabet size k.

It was shown in [4] this performance gap vanishes as $n \to \infty$. It may seem surprising given the $\Theta(\sqrt{k/n})$ minimax lower bound, but it is in fact reasonable: indeed, if the real distribution p is close to uniform, then the oracle would also have a hard time mapping the sorted probabilities to its original symbols, thereby incurring a large risk comparable to that incurred by our estimator. Hence, the gap between our estimator and the oracle's can still be controlled. A similar version of this question and its extensions was considered in [3] when the loss function is the KL divergence instead of the ℓ_1 loss.

3.1 Minimax risk for a given Θ_c

Before we discuss adaptive estimators, it is helpful for us to understand the minimax risk $R^*(\Theta_c)$ associated with a fixed parameter space Θ_c. It is given by the Pinsker theorem [6, Chapter 7].

Theorem 1 (Pinsker's Theorem)**.** *Let $\sigma_n^2 \triangleq \frac{\sigma^2}{n}$. For any $c > 0$,*

$$\lim_{n \to \infty} R^*(\Theta_c) = \frac{\sigma^2 c^2}{\sigma^2 + c^2}. \tag{12}$$

Pinsker's theorem gives us several interesting interpretations. When $c \to 0$, the risk approaches 0, which means we can just use 0 as an estimate of θ. On the other extreme, suppose $c \to \infty$. Then, the risk approaches σ^2 which matches Eq. (8), implying that the means estimator is minimax if we do not put a constraint on the parameter set.

Pinsker's Theorem has strong connections to our intuition from signal processing. Let's consider a Gaussian noise channel $Y = X + Z$ where the input X follows a Gaussian distribution $X \sim \mathcal{N}(0, P)$ and the noise Z follows a Gaussian distribution $Z \sim \mathcal{N}(0, N)$. Suppose we want to estimate X given Y. Then, in the Bayesian setting, it is well known that

$$\min_{\hat{X}(Y)} \mathbb{E}\left[\hat{X}(Y) - X\right]^2 = \frac{PN}{P+N}, \tag{13}$$

where equality is obtained at the conditional expecation of X given Y,

$$\hat{X}(Y) = \mathbb{E}[X|Y] = \frac{P}{P+N} \cdot Y. \tag{14}$$

To gain some intuition on how to interpret the $\frac{\sigma^2 c^2}{\sigma^2 + c^2}$ present in Eq. (12), we can think of each coordinate θ_i of θ as approximately $\theta_i^2 \approx \frac{c^2}{n}$, $\sigma_n^2 \approx \frac{\sigma^2}{n}$ and hence the corresponding risk in this coordinate is

$$\frac{\frac{\sigma^2}{n} \cdot \frac{c^2}{n}}{\frac{\sigma^2}{n} + \frac{c^2}{n}} = \frac{1}{n} \cdot \frac{\sigma^2 c^2}{\sigma^2 + c^2}. \tag{15}$$

Summing over all coordinates would yield $\frac{\sigma^2 c^2}{\sigma^2+c^2}$ as we wanted.

This hand-wavy argument describes the procedure used to prove Pinsker's Theorem—here we will provide a short sketch. Going back to our setting where the observations follow $X = \theta + \mathcal{N}\left(0, \frac{\sigma^2}{n} \cdot I_n\right)$, we would need to prove an upper bound and a lower bound that matches Eq. (12). The upper bound can be proved by constructing a coordinate-wise estimator $\hat{\theta}_c(i) = \frac{c^2 y_i}{\sigma^2+c^2}$ to show that the risk is upper bounded by $\frac{\sigma^2 c^2}{\sigma^2+c^2}$. For the lower bound, we can impose a prior $\theta_i \sim \mathcal{N}\left(0, \frac{c^2}{n}\right)$ and show that the Bayes risk is exactly $\frac{\sigma^2 c^2}{\sigma^2+c^2}$. Then, we can derive our lower bound by using the fact that minimax risk is lower bounded by Bayes risk, and the distribution of θ under the independent Gaussian prior *very tightly* concentration on Θ_c.

4 The James-Stein Estimator

Pinsker's Theorem (12) does not give us an explicit estimator for our adaptive estimation problem, since the estimators used in the analysis require knowledge of the parameter c of the true parameter space Θ_c. This issue can be resolved by using the James-Stein estimator [2]:

$$\hat{\theta}^{(JS)} \triangleq \left(1 - \frac{(n-2)\sigma_n^2}{\|Y\|_2^2}\right) \cdot Y. \tag{16}$$

As a historical note, the James-Stein estimator was not developed primarily for this problem. Instead, in the 1950's and 1960's Willard James and Charles Stein were working on problems regarding estimator admissibility. An estimator $\hat{\theta}^{(1)}$ is called *inadmissible* if there exists another estimator $\hat{\theta}^{(2)}$ such that

$$\mathbb{E}\ l(\hat{\theta}^{(1)}, \theta) \geq \mathbb{E}\ l(\hat{\theta}^{(2)}, \theta) \quad \text{for all } \theta \in \Theta, \tag{17}$$

and for at least one θ_0, the inequality is strict. In our problem setting, Stein tried to prove that the mean estimator $\hat{\theta} = Y$ (e.g., see Eq. (7)) is admissible, and was able to prove this for $n = 1$ and $n = 2$. However, he was not able to show that the mean estimator is admissible for $n = 3$, which led him to believe that the mean estimator is inadmissible. Eventually, he showed that the mean estimator is dominated by the James-Stein estimator $\hat{\theta}^{(JS)}$.

We will prove that when the true parameter space is Θ_c, the James-Stein estimator is able to achieve the optimal risk $\frac{\sigma^2 c^2}{\sigma^2+c^2}$ in Eq. (12). The proof will in part rely on the *Stein's Identity*:

Theorem 2 (Stein's Identity)**.** *Suppose $Z \sim \mathcal{N}(0,1)$, and let $f : \mathbb{R} \to \mathbb{R}$ be absolutely continuous. Then,*

$$\mathbb{E}\left[Z f(Z)\right] = \mathbb{E}[f'(Z)]. \tag{18}$$

Stein's identity can be used for proving the Central Limit Theorem with Stein's method, which we will not go into detail. It is natural to ask the question that goes in the other direction: If Eq. (20) is true for all absolutely continuous functions f, is Z standard Gaussian? Stein showed that this is true! In practice, sometimes checking Eq. (20) can be easier than calculating the characteristic function of Z. Stein's Identity can also be extended to the vector case.

Proof Since Z assumes a standard Gaussian distribution, its p.d.f.

$$\phi(z) \triangleq \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}).$$

By using "integration by parts", for any absolutely continuous function $f(z)$ we have

$$\mathbb{E}[f'(Z)] \triangleq \int_{-\infty}^{+\infty} f'(z)\phi(z)\,\mathrm{d}z \tag{19a}$$

$$= \int_{-\infty}^{+\infty} \phi(z)\,\mathrm{d}f(z) \tag{19b}$$

$$= \underbrace{\phi(z)f(z)\Big|_{-\infty}^{+\infty}}_{=0} - \int_{-\infty}^{+\infty} f(z) \underbrace{\mathrm{d}\phi(z)}_{=-z\phi(z)\,\mathrm{d}z} \tag{19c}$$

$$= \int_{-\infty}^{+\infty} f(z)z\phi(z)\,\mathrm{d}z \tag{19d}$$

$$= \mathbb{E}[Zf(Z)] \tag{19e}$$

□

Theorem 3 (Stein's Identity (vector version))**.** *Suppose $Z \sim \mathcal{N}(\theta, \sigma^2 I)$, and let $f : \mathbb{R}^n \to \mathbb{R}^n$ be absolutely continuous. Then,*

$$\mathbb{E}\left[Zf(Z)\right] = \mathbb{E}[\nabla f(Z)] = \frac{1}{\sigma^2}\mathbb{E}[(Z-\theta)f(Z)], \tag{20}$$

5 Stein Unbiased Risk Estimate

Suppose we observe a vector $Y \sim \mathcal{N}(\theta, \sigma_n^2 I)$, and $\hat{\theta}(Y)$ is an estimator of θ. Note that we cannot directly compute the risk because we do not know the real θ. As a result, we want to transform the risk expression into something that is computable, as to be shown below.

$$\mathbb{E}\|\theta - \hat{\theta}(Y)\|_2^2 = \mathbb{E}\|\theta - Y + Y - \hat{\theta}(Y)\|_2^2 \tag{21a}$$

$$= \mathbb{E}\|(\theta - Y) + (Y - \hat{\theta}(Y))\|_2^2 \tag{21b}$$

$$= \mathbb{E}\|\theta - Y\|_2^2 + \mathbb{E}\|Y - \hat{\theta}(Y)\|_2^2 - 2\mathbb{E}(Y-\theta)^T(Y - \hat{\theta}(Y)) \tag{21c}$$

$$= \mathbb{E}\|\theta - Y\|_2^2 + \mathbb{E}\|Y - \hat{\theta}(Y)\|_2^2 - 2\mathbb{E}(Y-\theta)^T(Y - \theta + \theta - \hat{\theta}(Y)) \tag{21d}$$

$$= \mathbb{E}\|\theta - Y\|_2^2 + \mathbb{E}\|Y - \hat{\theta}(Y)\|_2^2 - 2\mathbb{E}(Y-\theta)^T(Y - \theta) - 2\mathbb{E}(Y-\theta)^T(\theta - \hat{\theta}(Y)) \tag{21e}$$

$$= -\underbrace{\mathbb{E}\|\theta - Y\|_2^2}_{=n\sigma_n^2} + \underbrace{\mathbb{E}\|Y - \hat{\theta}(Y)\|_2^2}_{\text{observable}} - 2\underbrace{\mathbb{E}(Y-\theta)^T\theta}_{=0} + 2\mathbb{E}(Y-\theta)^T\hat{\theta}(Y) \tag{21f}$$

$$= -n\sigma_n^2 + \mathbb{E}\|Y - \hat{\theta}(Y)\|_2^2 + 2\sum_{i=1}^{n} \underbrace{\mathbb{E}\left[(Y_i - \theta_i)\hat{\theta}_i(Y)\right]}_{=\sigma_n^2\mathbb{E}\left[\frac{\partial\hat{\theta}_i(Y)}{\partial y_i}\right]} \tag{21g}$$

$$= -n\sigma_n^2 + \mathbb{E}\|Y - \hat{\theta}(Y)\|_2^2 + 2\sigma_n^2\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial\hat{\theta}_i(Y)}{\partial y_i}\right] \tag{21h}$$

$$= \mathbb{E}\hat{R} \quad \text{(unbiased estimation)} \tag{21i}$$

The James-Stein estimator (16) can be rewritten as,

$$\hat{\theta}^{(JS)} = Y + g(Y), \text{ where } g(Y) \triangleq -\frac{(n-2)\sigma_n^2}{\|Y\|_2^2}Y. \tag{22}$$

Plugging Eq. (22) into Eq. (21i), we get

$$\mathbb{E}\hat{R} = -n\sigma_n^2 + \mathbb{E}\left\|Y - \hat{\theta}(Y)\right\|_2^2 + 2\sigma_n^2 \sum_{i=1}^{n} \mathbb{E}\left[1 + \frac{\partial g_i}{\partial y_i}\right] \tag{23}$$

$$= n\sigma_n^2 + \mathbb{E}\left\|\frac{(n-2)\sigma_n^2}{\|Y\|_2^2} Y\right\|_2^2 + 2\sigma_n^2 \sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial g_i}{\partial y_i}\right] \tag{24}$$

$$= n\sigma_n^2 + \mathbb{E}\left\|\frac{(n-2)\sigma_n^2}{\|Y\|_2^2} Y\right\|_2^2 + 2\sigma_n^2 \sum_{i=1}^{n} \mathbb{E}\left[-(n-2)\sigma_n^2 \left(\frac{1}{\|Y\|_2^2} - \frac{2y_i^2}{\|Y\|_2^4}\right)\right] \tag{25}$$

$$= n\sigma_n^2 + \mathbb{E}\left[\frac{(n-2)^2\sigma_n^4}{\|Y\|_2^2}\right] + 2\sigma_n^2 \mathbb{E}\left[-\frac{(n-2)^2\sigma_n^2}{\|Y\|_2^2}\right] \tag{26}$$

$$= n\sigma_n^2 - \mathbb{E}\left[\frac{(n-2)^2\sigma_n^4}{\|Y\|_2^2}\right] \tag{27}$$

Now we can finally write out the expected risk of James-Stein estimator as follows

$$\mathbb{E}\hat{R}^{(JS)} = \mathbb{E}\left[n\sigma_n^2 - \frac{(n-2)^2\sigma_n^4}{\|Y\|_2^2}\right] \tag{28a}$$

$$= \underbrace{n\sigma_n^2}_{\text{MLE risk}} \underbrace{- (n-2)^2\sigma_n^4 \mathbb{E}\left[\frac{1}{\sum_i Y_i^2}\right]}_{\text{pointwise non-negative}}. \tag{28b}$$

From above we can see that the risk of James-Stein estimator will never exceed the risk given by the maximum likelihood estimator (MLE). However, there is a caveat: this computation only makes sense when $\mathbb{E}\left[\frac{1}{\sum_i Y_i^2}\right] < \infty$, and it is only true when $n > 2$. For further simplifying the expectation term in Eq. (28b), we will need to use the following lemma to deal with non-central χ^2 distribution.

Lemma 4. *Let random variable w assume a non-central χ^2 distribution with n degrees of freedom, i.e. $w = \sum_{i=1}^{n} \theta_i^2$, where $\theta_i \sim \mathcal{N}(m_i, 1)$ and $\sum_i m_i^2 = \delta$, then $w \sim \chi^2_{n+2k}$ where $k \sim Poi(\delta/2)$.*

Note that the ith element of the multivariate Y variable can be reparametrized in the following form,

$$Y_i = \sigma_n \left(\frac{\theta_i}{\sigma_n} + z_i\right), \text{ where } z_i \sim \mathcal{N}(0, 1). \tag{29}$$

Then we have,

$$\|Y\|_2^2 = \sum_{i=1}^{n} Y_i^2 = \sigma_n^2 W, \tag{30}$$

where random variable W assumes a non-central χ^2 distribution with $\delta = \sum_{i=1}^{n} \frac{\theta_i^2}{\sigma_n^2}$.

By using Lemma 4, we can now further simplify the expectation term in Eq. (28b). Note that below we

use the properties of inverse χ^2 distribution and Jensen's inequality.

$$\mathbb{E}\left[\frac{1}{\sum_i Y_i^2}\right] = \frac{1}{\sigma_n^2}\mathbb{E}\left[\mathbb{E}\left[\frac{1}{\chi^2_{n+2k}}\,\middle|\, k\right]\right] \tag{31a}$$

$$= \frac{1}{\sigma_n^2}\mathbb{E}\left[\frac{1}{n+2k-2}\right] \tag{31b}$$

$$\geq \frac{1}{\sigma_n^2}\left(\frac{1}{n-2+2\mathbb{E}k}\right) \tag{31c}$$

$$= \frac{1}{(n-2)\sigma_n^2 + \sum_{i=1}^n \theta_i^2} \tag{31d}$$

$$= \frac{1}{(n-2)\sigma_n^2 + \|\theta\|_2^2} \tag{31e}$$

If we plug Eq. (31e) back into Eq. (28b) and rearrange the terms, we can get the expression for the expectation of risk as below.

$$\hat{R}^{(JS)} \leq 2\sigma_n^2 + \frac{(n-2)\sigma_n^2 \cdot \|\theta\|_2^2}{(n-2)\sigma_n^2 + \|\theta\|_2^2} \tag{32}$$

$$= \frac{2}{n}\sigma^2 + \frac{\sigma^2 \cdot \|\theta\|_2^2}{\sigma^2 + \frac{n}{n-2}\|\theta\|_2^2} \tag{33}$$

$$\to \frac{\sigma^2 \cdot \|\theta\|_2^2}{\sigma^2 + \|\theta\|_2^2} \text{ as } n \to \infty. \tag{34}$$

From above we can immediately see its relationship with Pinsker's Theorem (12), which reveals the underlying link between James-Stein estimator and adaptive estimation.

References

[1] Y. Han, J. Jiao, and T. Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354, 2015.

[2] W. James and C. Stein. Estimation with quadratic loss. In *Breakthroughs in statistics*, pages 443–460. Springer, 1992.

[3] A. Orlitsky and A. T. Suresh. Competitive distribution estimation: Why is good-turing good. In *Advances in Neural Information Processing Systems*, pages 2143–2151, 2015.

[4] G. Valiant and P. Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 142–155. ACM, 2016.

[5] M. J. Wainwright. Constrained forms of statistical minimax: Computation, communication and privacy.

[6] L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.