LECTURE 5 NOTES

1. Bayesian point estimators. In the conventional (frequentist) approach to statistical inference, the parameter $\theta \in \Theta$ is considered a fixed quantity. In the Bayesian approach, it is considered stochastic. Its distribution over Θ before observing any data, called the *prior*, reflects the investigator's prior beliefs about θ. After obtaining a sample $\mathbf{x}$, the investigator updates his or her beliefs by Bayes' rule, hence the name.

Let $\pi(\theta)$ be the prior density on Θ and $f_\theta(x)$ be a set of densities on $\mathcal{X}$. The distribution of $\boldsymbol{\theta} \mid \mathbf{x}$ is called the *posterior distribution* and is given by Bayes' rule:

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{f_\pi(x)},$$

where $f_\pi(x) := \int_\Theta f(x \mid \theta)\pi(\theta)d\theta$ is the marginal density of $\mathbf{x}$. It reflects the investigator's updated beliefs about $\boldsymbol{\theta}$ after observing $\mathbf{x}$. Its expected value is often used as a point estimate of $\boldsymbol{\theta}$.

EXAMPLE 1.1. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$. *A Bayesian investigator assumes a* $\text{beta}(a, b)$ *prior on* $\mathbf{p}$. *It is possible to show that the joint density of* $\mathbf{t} = \sum_{i\in[n]} \mathbf{x}_i$ *and* $\mathbf{p}$ *is*

$$f(t, p) = \binom{n}{t} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{t+a-1}(1-p)^{n-t+b-1}.$$

The marginal density of $\mathbf{t}$ *is*

$$f_{\text{beta}(a,b)}(t) = \binom{n}{t} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(t+a)\Gamma(n-t+b)}{\Gamma(n+a+b)},$$

and the posterior density of p *is*

$$\pi(p \mid t) = \frac{\Gamma(n+a+b)}{\Gamma(t+a)\Gamma(n-t+b)} p^{t+a-1}(1-p)^{n-t+b-1},$$

which is the density of a $\text{beta}(t+a, n-t+b)$ *random variable. The Bayes estimator is*

$$\hat{p}_B := \mathbf{E}_\pi[\mathbf{p} \mid \mathbf{t}] = \frac{\mathbf{t}+a}{a+b+n}$$

The Bayes estimator $\hat{p}_B$ *has the form*

$$\hat{p}_B = \frac{n}{a+b+n}\frac{\mathbf{t}}{n} + \frac{a+b}{a+b+n}\frac{a}{a+b}. \tag{1.1}$$

Thus it is a convex combination of the sample mean $\frac{\mathbf{t}}{n}$ *and the prior mean* $\frac{a}{a+b}$*. As the sample size grows,* $\hat{p}_B$ *converges to* $\frac{\mathbf{t}}{n}$*.*

EXAMPLE 1.2. *Let* $\mathbf{x} \sim \mathcal{N}(\mu, 1)$*. A Bayesian investigator assumes a* $\mathcal{N}(a, b^2)$ *prior on* μ*. The posterior density of* $\boldsymbol{\mu}$ *after observing* $\mathbf{x}$ *is*

$$\pi(\mu \mid \mathbf{x}) \propto \underbrace{\exp\left(-\tfrac{1}{2}(x-\mu)^2\right)}_{f(x|\mu)} \underbrace{\exp\left(-\tfrac{(\mu-a)^2}{2b^2}\right)}_{\pi(\mu)}$$

It is possible to show that the posterior is Gaussian:

$$\pi(\mu \mid \mathbf{x}) \propto \exp\left(-\frac{(\mu-\tilde{a})^2}{2\tilde{b}^2}\right),$$

where

$$\tilde{b}^2 = \left(1 + \frac{1}{b^2}\right)^{-1/2}, \; \tilde{a} = \frac{\mathbf{x} + \frac{a}{b^2}}{1 + \frac{1}{b^2}} = \frac{b^2\mathbf{x} + a}{1 + b^2}.$$

The Bayes estimator is the posterior mean, which is also a convex combination of the sample $\mathbf{x}$ *and the prior mean* a*.*

In practice, it is usually not possible to derive the expected value of the posterior in closed-form, and we must evaluate Bayes estimators by *Monte Carlo methods*.

In the preceding two examples, we chose the expected value of the posterior as the point estimator. Another option is the mode of the posterior, also known as the *maximum a posteriori* (MAP) estimator:

$$\hat{\theta}_{\text{MAP}} \in \arg\max_{\theta \in \Theta} \pi(\theta \mid \mathbf{x}) = \arg\max_{\theta \in \Theta} f(\mathbf{x} \mid \theta)\pi(\theta). \tag{1.2}$$

The main advantage of MAP estimators over Bayes estimators is they are usually easier to evaluate that Bayes estimators. For this reason, MAP estimators are sometimes known as the poor man's Bayes estimator.

EXAMPLE 1.3. *Consider the Bayesian linear model*

$$\boldsymbol{\beta} \sim \mathcal{N}(0, \lambda^{-1} I_p)$$
$$\mathbf{y}_i \mid \{\mathbf{x}_i, \boldsymbol{\beta}\} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\beta}, 1).$$

We remark that the task is to learn the distribution of $\mathbf{y}_i \mid \mathbf{x}_i$*: the distribution of* $\mathbf{x}_i$ *is irrelevant. The posterior is*

$$\pi(\beta \mid X, y) \propto \prod_{i \in [n]} \exp\Big(-\tfrac{1}{2}\|y_i - x_i^T\beta\|_2^2\Big) \exp\left(-\tfrac{\lambda}{2}\|\theta\|_2^2\right).$$

Maximizing the posterior is equivalent to maximizing

$$\log \pi(\beta \mid X, y) = -\frac{1}{2} \sum_{i \in [n]} \|y_i - x_i^T \beta\|_2^2 - \frac{\lambda}{2} \|\theta\|_2^2,$$

which is akin to ridge regression.

We remark that MAP estimators are generally interpretable as regularized MLE's: the regularizer is the log-density of the prior. Another prominent example is the lasso estimator:

$$\hat{\beta}_{\text{lasso}} \in \arg\min_{\beta \in \mathbf{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1,$$

which is the MAP estimator under a double-exponential prior.

A subtle drawback of the MAP estimator is it is not equivariant under reparametrization. Let $\pi(\theta \mid \mathbf{x})$ be the posterior density of $\boldsymbol{\theta}$ and $\hat{\theta}_{\text{MAP}}$ be its mode. The posterior density of $g(\theta)$, where $g : \Theta \to \mathcal{H}$ is invertible, is

$$\pi'(\eta \mid x) = \pi(g^{-1}(\eta) \mid x) \left| \nabla g^{-1}(\eta) \right|.$$

Due to the presence of the Jacobian term, the maximizer of $\pi'(\eta \mid x)$ is generally not $g(\hat{\theta}_{\text{MAP}})$.

EXAMPLE 1.4. *Consider a Bayesian investigation of the expected value of a* $\text{Ber}(p)$ *random variable. The investigator assumes a uniform prior on* $[0, 1]$*. Before collecting any data, the posterior is just the prior, and a MAP estimator of* p *is any point in* $[0, 1]$*. However, the prior of* $q = \sqrt{p}$ *is*

$$\pi'(q) = \pi(q^2) \, |2q| = \mathbf{1}_{[0,1]}\big(q^2\big) 2q,$$

which is maximized at 1. The prior of $r = 1 - \sqrt{1-p}$ *is*

$$\pi''(r) = \pi\big((1-r)^2\big) \, |2(1-r)| \,,$$

which is maximized at 0.

A key design choice in a Bayesian investigation is the choice of prior. If prior knowledge is available, it is often desirable to incorporate this knowledge into the choice of prior. We summarize some popular approaches to choosing a prior.

1. **empirical Bayes:** estimate (the parameters of) the prior from data. The marginal distribution of $\mathbf{x}$ usually depends on the parameters of the prior, which can be used to estimate the parameters. We give an example of an empirical Bayes estimator later in the course.

2. **hierarchical Bayes:** impose a *hyper prior* on the parameters of the prior. The choice of hyper prior usually has a smaller impact on the Bayes estimator than the choice of prior.
3. **robust Bayes:** look for estimators that perform well under all priors in some family of priors.

Bayes estimators are less popular in practice than maximum likelihood or MoM estimators because the choice of prior is usually subjective. Two investigators given the same dataset may arrive at different estimates of a parameter due to differences in their priors. Thus, unlike the MLE or MoM estimators, Bayes estimators are inherently subjective.

2. Evaluating point estimators.

DEFINITION 2.1. *The mean squared error (MSE) of an estimator $\delta(\mathbf{x})$ of a parameter θ is* $\mathbf{E}_\theta\Big[\|\delta(\mathbf{x}) - \theta\|_2^2\Big]$.[1]

The MSE is the average discrepancy between the estimator $\delta(\mathbf{x})$ and the unknown parameter θ in the ℓ_2 norm. For any estimator, the MSE decomposes into *bias* and *variance* terms:

$$\begin{aligned}
\mathbf{E}_\theta\Big[\|\delta(\mathbf{x}) - \theta\|_2^2\Big] &= \mathbf{E}_\theta\Big[\|\delta(\mathbf{x}) - \mathbf{E}_\theta\big[\delta(\mathbf{x})\big] + \mathbf{E}_\theta\big[\delta(\mathbf{x})\big] - \theta\|_2^2\Big] \\
&= \mathbf{E}_\theta\Big[\|\delta(\mathbf{x}) - \mathbf{E}_\theta\big[\delta(\mathbf{x})\big]\|_2^2\Big] + \mathbf{E}_\theta\Big[\|\,\mathbf{E}_\theta\big[\delta(\mathbf{x})\big] - \theta\|_2^2\Big] \\
&\quad + 2\,\mathbf{E}_\theta\Big[\big(\delta(\mathbf{x}) - \mathbf{E}_\theta\big[\delta(\mathbf{x})\big]\big)^T\big(\mathbf{E}_\theta\big[\delta(\mathbf{x})\big] - \theta\big)\Big] \\
&= \underbrace{\mathbf{E}_\theta\Big[\|\delta(\mathbf{x}) - \mathbf{E}_\theta\big[\delta(\mathbf{x})\big]\|_2^2\Big]}_{\text{variance}} + \underbrace{\|\,\mathbf{E}_\theta\big[\delta(\mathbf{x})\big] - \theta\|_2^2}_{\text{bias}^2}.
\end{aligned} \tag{2.1}$$

The bias term is the difference between the expected value of the estimator and the target and is a measure of the estimator's accuracy. An estimator whose bias vanishes for any $\theta \in \Theta$ is *unbiased*.

MSE is by no means the only error metric that practitioners consider. It is a special case of a risk function. The study of the performance of estimators evaluated by risk functions is a branch of *decision theory*.

Decision theory formalizes a statistical investigations as a decision problem. After observing $\mathbf{x} \sim F \in \mathcal{F}$, the investigator "makes a decision" regarding the unknown parameter $\theta \in \Theta$. The set of allowed decisions is called the *action space* $\mathcal{A}$. In point estimation, the decision is typically the point

[1] The subscript on $\mathbf{E}$ means the expectation is taken with respect to $F_\theta \in \mathcal{F}$.

estimate. Thus $\mathcal{A}$ is often just Θ. After a decision is made, the investigator incurrs a loss given by a loss function $l : \Theta \times \mathcal{A} \to \mathbf{R}_+$. By convention, "bad" decisions incur higher losses, and the investigator seeks to minimize his or her losses. Some examples of loss functions are

1. *square loss*: $l(\theta, a) = \|\theta - a\|_2^2$,
2. *absolute value loss*: $l(\theta, a) = \|\theta - a\|_1$,
3. *zero-one-loss*: $l(\theta, a) = 1 - \mathbf{1}_\theta(a)$,
4. *log loss*: $l(\theta, a) = \frac{f_\theta(\mathbf{x})}{f_a(\mathbf{x})}$.

The performance of a *decision rule* $\delta : \mathcal{X} \to \mathcal{A}$ is summarized by its *risk function*:

$$\text{Risk}_\delta(\theta) = \mathbf{E}_\theta[l(\theta, \delta(\mathbf{x}))].$$

The MSE of a point estimator $\delta(\mathbf{x})$ is the risk function of the decision rule δ under the square loss function $l(\theta, a) = \|\theta - a\|_2^2$.

2.1. *Rao-Blackwellization.* As we shall see, there is an intimate connection between data reduction and good point estimators (in the decision theoretic sense). When the loss function is convex (in δ), it is always possible to reduce the risk of an estimator by *Rao-Blackwellization*.

THEOREM 2.2 (Rao-Blackwell). *Let* $\mathbf{t}$ *be a sufficient statistic for model* $\mathcal{F} := \{F_\theta : \theta \in \Theta\}$. *For any loss function* $l : \Theta \times \mathcal{A} \to \mathbf{R}_+$ *that is convex in its second argument, we have*

$$\mathbf{E}_\theta\Big[l\left(\theta, \mathbf{E}\big[\delta(\mathbf{x}) \mid \mathbf{t}\big]\right)\Big] \leq \mathbf{E}_\theta\big[l(\theta, \delta(\mathbf{x}))\big] \text{ for any } \theta \in \Theta.$$

Further, if l *is strictly convex in its second argument, the inequality is strict unless* $\delta(\mathbf{x}) \overset{a.s.}{=} \mathbf{E}\big[\delta(\mathbf{x}) \mid \mathbf{t}\big]$.

PROOF. By the tower property of conditional expectations,

$$\mathbf{E}_\theta\big[l(\theta, \delta(\mathbf{x}))\big] = \mathbf{E}_\theta\Big[\mathbf{E}\big[l(\theta, \delta(\mathbf{x})) \mid \mathbf{t}\big]\Big],$$

which, by Jensen's inequality, is at least

$$\geq \mathbf{E}_\theta\Big[l\left(\theta, \mathbf{E}\big[\delta(\mathbf{x}) \mid \mathbf{t}\big]\right)\Big].$$

If l is strictly convex in its second argument, unless $\delta(\mathbf{x}) \overset{a.s.}{=} \mathbf{E}\big[\delta(\mathbf{x}) \mid \mathbf{t}\big]$, Jensen's inequality is strict . □

The Rao-Blackwell theorem shows that it is possible to reduce the risk of any point estimator by conditioning on a sufficient statistic. The estimator $\delta_{\mathrm{RB}}(\mathbf{t}) := \mathbf{E}\big[\delta(\mathbf{x}) \mid \mathbf{t}\big]$ is sometimes called a Rao-Blackwellized estimator. By modifying the proof of the Rao-Blackwell theorem, it is possible to show that if $\mathbf{t}$ is a minimal sufficient statistic and $\mathbf{t}'$ is another sufficient statistic, then

$$\mathbf{E}_\theta\Big[l\left(\theta, \mathbf{E}\big[\delta(\mathbf{x}) \mid \mathbf{t}\big]\right)\Big] \leq \mathbf{E}_\theta\big[l\left(\theta, \mathbf{E}\big[\delta(\mathbf{x}) \mid \mathbf{t}'\big]\right)\Big] \text{ for any } \theta \in \Theta.$$

The astute reader may observe that the proof of the Rao-Blackwell theorem is valid whether $\mathbf{t}$ is a sufficient statistic or not. Thus conditioning on any statistic reduces the risk. Although this is true, the law of $\delta(\mathbf{x}) \mid \mathbf{t}$ generally depends on the unknown parameter θ. Thus the Rao-Blackwellized "estimator" is, in fact, not an estimator.

2.2. *Admissibility.* Recall the risk of an estimator δ

$$R_\delta(\theta) = \mathbf{E}_\theta\big[l(\theta, \delta(\mathbf{x}))\big],$$

which led us to study risk-study notions of optimality for estimators. Unfortunately, there is usually no uniformly optimal estimator.

EXAMPLE 2.3. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)$. *The MLE of* p *is* $\hat{p}_{\mathrm{ML}} = \bar{\mathbf{x}}$, *and the Bayes estimator is*

$$\hat{p}_B = \mathbf{E}_\pi[\mathbf{p} \mid \mathbf{t}] = \frac{n\bar{\mathbf{x}} + a}{a + b + n} \tag{2.2}$$

Since the MLE is unbiased, its MSE is its variance, which is $\frac{p(1-p)}{n}$. *The MSE of the Bayes estimator is*

$$\begin{aligned}
\mathrm{MSE}_{\hat{p}_B}(p) &= \mathbf{var}_p\big[\hat{p}_B\big] + \big(\mathbf{E}_p\big[\hat{p}_B\big] - p\big)^2 \\
&= \mathbf{var}_p\left[\frac{n\bar{\mathbf{x}} + a}{a + b + n}\right] + \left(\mathbf{E}_p\left[\frac{n\bar{\mathbf{x}} + a}{a + b + n}\right] - p\right)^2 \\
&= \frac{np(1-p)}{(a+b+n)^2} + \left(\frac{np + a}{a + b + n} - p\right)^2,
\end{aligned}$$

which is a quadratic function of p. *It is possible to show that by choosing* $a = b = \sqrt{\frac{n}{4}}$, *the MSE is constant in* p:

$$\mathrm{MSE}_{\hat{p}_B}(p) = \frac{n}{4(n + \sqrt{n})^2}.$$

We observe that the MSE of the MLE is smaller than that of $\hat{p}_B$ *when* p *is near 0 or 1, but larger when* p *is near* $\frac{1}{2}$. *Thus neither estimator dominates the other uniformly on* $p \in [0, 1]$.

Although there is usually no uniformly optimal estimator, there are uniformly suboptimal estimators.

DEFINITION 2.4. *An decision rule δ is* inadmissible *if there is another decision rule δ' such that $R_{\delta'}(\theta) \leq R_\delta(\theta)$ for any $\theta \in \Theta$ and $R_{\delta'}(\theta) < R_\delta(\theta)$ at some $\theta \in \Theta$. Otherwise, δ is* admissible.

Intuitively, an inadmissible estimator is uniformly bested by another estimator, so, from a decision theoretic perspective, there is no good reason to use inadmissible estimators. We remark that although inadmissible estimators are "bad", admissible estimators are not necessarily "good". An estimator is admissible if it has (strictly) smaller risk than any other estimator at a single $\theta \in \Theta$. Its risk at any other θ may be arbitrarily bad.

3. Bayes estimators.

DEFINITION 3.1. *Let π be a prior on the parameter space Θ. The* Bayes risk *of a decision rule δ is*

$$\mathbf{E}_\pi\big[R_\delta(\boldsymbol{\theta})\big] = \int_\Theta R_\delta(\theta)\pi(\theta)d\theta.$$

It is possible to minimize the Bayes risk to derive a *Bayes estimator*. We begin by noticing the Bayes risk has the form

$$\begin{aligned}\mathbf{E}_\pi\big[R_\delta(\boldsymbol{\theta})\big] &= \int_{\Theta\times\mathcal{X}} l(\theta,\delta(x))f(x \mid \theta)\pi(\theta)dxd\theta \\ &= \int_{\Theta\times\mathcal{X}} l(\theta,\delta(x))\pi(\theta \mid x)f_\pi(x)dxd\theta \\ &= \int_{\mathcal{X}} \mathrm{PostRisk}_\delta(x)f_\pi(x)dx.\end{aligned}$$

where

$$\mathrm{PostRisk}_\delta(x) = \mathbf{E}_{\boldsymbol{\theta}\sim\pi(\cdot|x)}\big[l(\boldsymbol{\theta},\delta(x))\big] = \int_\Theta l(\theta,\delta(x))\pi(\theta \mid x)d\theta$$

is the *posterior risk* of the estimator δ. We observe that the posterior risk does not depend on θ; it only depends only on x. By choosing $\delta(x)$ to minimize the posterior risk, we minimize the Bayes risk. In practice, it is only necessary to minimize the posterior at the observed $\mathbf{x}$. That is,

$$\delta_B(x) := \arg\min_{a\in\mathcal{A}} \mathbf{E}_{\boldsymbol{\theta}\sim\pi(\cdot|x)}\big[l(\boldsymbol{\theta},a)\big].$$

EXAMPLE 3.2. *The posterior MSE of an estimator δ is*

$$\begin{aligned}&\operatorname{PostMSE}_\delta(x)\\&\quad= \mathbf{E}_{\boldsymbol{\theta}\sim\pi(\cdot\,|\,x)}\big[\|\boldsymbol{\theta}-\delta(x)\|_2^2\big]\\&\quad= \mathbf{E}_{\boldsymbol{\theta}\sim\pi(\cdot\,|\,x)}\big[\|\boldsymbol{\theta}\|_2^2\big]+2\,\mathbf{E}_{\boldsymbol{\theta}\sim\pi(\cdot\,|\,x)}\big[\boldsymbol{\theta}\big]^T\delta(x)+\|\delta(x)\|_2^2,\end{aligned}$$

which is a quadratic function of $\delta(x)$. It is minimized at

$$\delta_B(x)=\mathbf{E}_{\boldsymbol{\theta}\sim\pi(\cdot\,|\,x)}\big[\boldsymbol{\theta}\big].$$

Thus the Bayes estimator is the expected value of the posterior. Example 2.3 is a concrete example of a Bayes estimator.

EXAMPLE 3.3. *Instead of the square loss function, consider the zero-one loss function*

$$l(\theta,\delta)=1-\mathbf{1}_\theta(\delta)=\begin{cases}0 & \theta=\delta\\ 1 & \textit{otherwise}\end{cases}.$$

If Θ is finite, the posterior risk is

$$\operatorname{PostRisk}_\delta(x)=\sum_{\theta\in\Theta}l(\theta,\delta(x))\pi(\theta\mid x)=1-\pi(\delta(x)\mid x).$$

To minimize the posterior risk, we should choose $\delta(x)$ so that $\pi(\delta(x)\mid x)$ is as large as possible. Thus the Bayes estimator under the zero-one loss function is the MAP estimator.

When minimization of the posterior risk cannot be done analytically, it is often done numerically:

$$\delta_B(x)\in\operatorname*{arg\,min}_{\delta\in\Theta}\operatorname{PostRisk}_\delta(x)=\operatorname*{arg\,min}_{\delta\in\Theta}\log\big(\operatorname{PostRisk}_\delta(x)\big).$$

We remark that the problem is similar to evaluating the MLE. Thus most numerical methods for evaluating the MLE are directly applicable to minimizing the posterior risk.

Before moving on to minimax estimators, we comment on the admissibility of Bayes estimators.

THEOREM 3.4. *A unique Bayes estimator is admissible.*

PROOF. If there is another estimator θ such that $R_\delta(\theta)\le R_{\delta_B}(\theta)$ for any $\theta\in\Theta$ and $R_\delta(\theta)<R_{\delta_B}(\theta)$ at some θ, then

$$\mathbf{E}_\pi\big[R_\delta(\boldsymbol{\theta})\big]\le\mathbf{E}_\pi\big[R_{\delta_B}(\boldsymbol{\theta})\big]. \tag{3.1}$$

Thus $\delta(\mathbf{x})$ is Bayes. Since δ_B is unique, $\delta_B(x)=\delta(x)$ for all $x\in\mathcal{X}$. □

Yuekai Sun
Berkeley, California
November 23, 2015