A Primer of Frequentist and Bayesian Inference in Inverse Problems

Philip B Stark[1] & Luis Tenorio[2]

[1]University of California at Berkeley
[2]Colorado School of Mines

0.1 Introduction

Inverse problems seek to learn about the world from indirect, noisy data. They can be cast as statistical estimation problems and studied using statistical decision theory, a framework that subsumes Bayesian and frequentist methods. Both Bayesian and frequentist methods require a stochastic model for the data, and both can incorporate constraints on the possible states of the world. Bayesian methods require that constraints be re-formulated as a *prior probability distribution*, a stochastic model for the unknown state of the world. If the state of the world is a random variable with a known distribution, Bayesian methods are in some sense optimal. Frequentist methods are easier to justify when the state of the world is unknown but not necessarily random, or random but with an unknown distribution; some frequentist methods are then optimal—in other senses.

Parameters are numerical properties of the state of the world. *Estimators* are quantities that can be computed from the data without knowing the state of the world. Estimators can be compared using *risk functions* which quantify the expected "cost"

Computational Methods for Large-Scale Inverse Problems and Quantification of Uncertainty. Edited by ...

of using a particular estimator when the world is in a given state. (The definition of cost should be dictated by the scientific context, but some definitions, such as mean squared error, are used because they lead to tractable computations.)

Generally, no estimator has the smallest risk for all possible states of the world: there are tradeoffs, which Bayesian and frequentist methods address differently. Bayesian methods seek to minimize the expected risk when the state of the world is drawn at random according to the prior probability distribution. One frequentist approach—*minimax estimation*—seeks to minimize the maximum risk over all states of the world that satisfy the constraints.

The performance of frequentist and Bayesian estimators can be compared both with and without a prior probability distribution for the state of the world. Bayesian estimators can be evaluated from a frequentist perspective, and vice versa. Comparing the minimax risk with the Bayes risk measures how much information the prior probability distribution adds that is not present in the constraints or the data.

This chapter sketches frequentist and Bayesian approaches to estimation and inference, including some differences and connections. The treatment is expository, not rigorous. We illustrate the approaches with two examples: a concrete one-dimensional problem (estimating the mean of a Normal distribution when that mean is known to lie in the interval $[-\tau, \tau]$), and an abstract linear inverse problem.

For a more philosophical perspective on the frequentist and Bayesian interpretations of probability and models, see Freedman & Stark (2003); Freedman (1995). For more careful treatments of the technical aspects, see Berger (1985); Evans & Stark (2002); Le Cam (1986).

0.2 Prior Information and Parameters: What do you know, and what do you want to know?

This section lays out some of the basic terms, most of which are shared by frequentist and Bayesian methods. Both schools seek to learn about the state of the world—to estimate parameters—from noisy data. Both consider constraints on the possible states of the world, and both require a stochastic measurement model for the data.

0.2.1 The state of the world, measurement model, parameters and likelihoods

The term "model" is used in many ways in different communities. In the interest of clarity, we distinguish among three things sometimes called "models," namely, the state of the world, the measurement model, and parameters.

The the state of the world, denoted by θ, is a mathematical representation of the physical system of interest; for example, a parametrized representation of seismic velocity as a function of position in the Earth, of the angular velocity of material in the Sun, or of the temperature of the cosmic microwave background radiation as a function of direction. Often in physical problems, some states of the world can be

ruled out by physical theory or prior experiment. For instance, mass densities and energies are necessarily nonnegative. Transmission coefficients are between 0 and 1. Particle velocities are less than the speed of light. The rest mass of the energy stored in Earth's magnetic field is less than the mass of the Earth (Backus 1989). The set Θ will represent the possible states of the world (values of θ) that satisfy the constraints. That is, we know *a priori* that $\theta \in \Theta$.[1]

The observations Y are related to the particular state of the world θ through a *measurement model* that relates the probability distribution of the observations to θ. The set of possible observations is denoted $\mathcal{Y}$, called the *sample space*. Typically, Y is an n-dimensional vector of real numbers; then, $\mathcal{Y}$ is $\mathbb{R}^n$. Depending on θ, Y is more likely to take some values in $\mathcal{Y}$ than others. The probability distribution of the data Y when the state of the world is η (i.e., when $\theta = \eta$) is denoted $\mathbb{P}_\eta$; we write $Y \sim \mathbb{P}_\eta$. (The "true" state of the world is some particular—but unknown—$\theta \in \Theta$; η is a generic element of Θ that might or might not be equal to θ.) We shall assume that the set of distributions $\mathcal{P} \equiv \{\mathbb{P}_\eta : \eta \in \Theta\}$ is dominated by a common measure μ.[2] (In the special case that μ is Lebesgue measure, that just means that all the probability distributions $\mathcal{P} \equiv \{\mathbb{P}_\eta : \eta \in \Theta\}$ have densities in the ordinary sense.) The *density of $\mathbb{P}_\eta$ (with respect to μ) at y* is

$$p_\eta(y) \equiv d\mathbb{P}_\eta / d\mu|_y. \tag{1}$$

The *likelihood of η given $Y = y$* is $p_\eta(y)$ viewed as a function of η, with y fixed.

For example, suppose that for the purposes of our experiment, the state of the world can be described by a single number $\theta \in \mathbb{R}$ that is known to be in the interval $[-\tau, \tau]$, and that our experiment measures θ with additive Gaussian noise that has mean zero and variance 1. Then the measurement model is

$$Y = \theta + Z, \tag{2}$$

where Z is a standard Gaussian random variable (we write $Z \sim N(0,1)$). The set $\Theta = [-\tau, \tau]$. Equivalently, we may write $Y \sim N(\theta, 1)$ with $\theta \in [-\tau, \tau]$. (The symbol $\sim$ means "has the probability distribution" or "has the same probability distribution as.") Thus $\mathbb{P}_\eta$ is the Gaussian distribution with mean η and variance 1. This is called the *bounded normal mean* (BNM) problem. The BNM problem is of theoretical interest, and is a building block for the study of more complicated problems in higher dimensions; see, for example, Donoho (1994). The dominating measure μ in this problem can be taken to be Lebesgue measure. Then the density of $\mathbb{P}_\theta$ at y is

$$\phi_\theta(y) \equiv \frac{1}{\sqrt{2\pi}} e^{-(y-\theta)^2/2}, \tag{3}$$

[1] There is a difference between the physical state of the world and a numerical discretization or approximation of the state of the world for computational convenience. The numerical approximation to the underlying physical process contributes uncertainty that is generally ignored. For discussion, see, e.g., Stark (1992b).

[2] By defining μ suitably, this can allow us to work with "densities" even if the family $\mathcal{P}$ contains measures that assign positive probability to individual points. Assuming that there is a dominating measure is a technical convenience that permits a general definition of likelihoods.

and the likelihood of η given $Y = y$ is $\phi_\eta(y)$ viewed as a function of η with y fixed.

As a more general example, consider the following canonical linear inverse problem. The set Θ is a norm or semi-norm ball in a separable, infinite-dimensional Banach space (for example, Θ might be a set of functions whose integrated squared second derivative is less than some constant $C < \infty$, or a set of nonnegative functions that are continuous and bounded).[3] The data Y are related to the state of the world θ through the action of a linear operator K from Θ to $\mathcal{Y} = \mathbb{R}^n$, with additive noise:

$$Y = K\theta + \epsilon, \quad \theta \in \Theta, \tag{4}$$

where the probability distribution of the noise vector ϵ is known. We assume that $K\eta = (K_1\eta, K_2\eta, \ldots, K_n\eta)$ for $\eta \in \Theta$, where $\{K_j\}_{j=1}^n$ are linearly independent bounded linear functionals on Θ. Let $f(y)$ denote the density of ϵ with respect to a dominating measure μ. Then $p_\eta(y) = f(y - K\eta)$. The BNM problem is an example of a linear inverse problem with K the identity operator, $\mathcal{Y} = \mathbb{R}$, $\Theta \equiv [-\tau, \tau]$, $\epsilon \sim N(0, 1)$, $f(y) = \phi_0(y)$.

A *parameter* $\lambda = \lambda[\theta]$ is a property of the state of the world. The entire description of the state of the world, θ, could be considered to be a parameter; then λ is the identity operator. Alternatively, we might be interested in a simpler property of θ. For example, in gravimetry, the state of the world θ might be mass density as a function of position in Earth's interior, and the parameter of interest, $\lambda[\theta]$, might be the average mass density in some region below the surface. In that case, the rest of θ is a *nuisance parameter*: it can affect the (probability distribution of the) measurements, but it is not of primary interest.

Our lead examples in this paper are the "bounded normal mean" problem and the canonical linear inverse problem just described.

0.2.2 Prior and posterior probability distributions

In the present framework, there is prior information about the state of the world θ expressed as a constraint $\theta \in \Theta$. Frequentists use such constraints as-is. Bayesians augment the constraints using *prior probability distributions*. Bayesians treat the value of θ as a realization of a random variable that takes values in Θ according to a prior probability distribution π.[4] The constraint $\theta \in \Theta$ is reflected in the fact that π assigns probability 1 (or at least high probability) to the set Θ.[5] In the BNM problem,

[3] Separable Banach spaces are measurable with respect to the σ-algebra induced by the norm topology, a fact that ensures that prior probability distributions—required by Bayesian methods—can be defined.

[4] To use prior probability distributions, Θ must be a measurable space; frequentist methods generally do not need the set of parameters to be measurable. We will not worry about technical details here. For rigor, see Le Cam (1986).

[5] Freedman (1995) writes, "My own experience suggests that neither decision-makers nor their statisticians do in fact have prior probabilities. A large part of Bayesian statistics is about what you would do if you had a prior. For the rest, statisticians make up priors that are mathematically convenient or attractive. Once used, priors become familiar; therefore, they come to be accepted as 'natural' and are liable to be used again; such priors may eventually generate their own technical literature." And, "Similarly, a large part of [frequentist] statistics is about what you would do if you had a model; and all of us spend

a Bayesian would use a prior probability distribution π that assigns probability 1 to the interval $[-\tau, \tau]$. For instance, she might use as the prior π the "uninformative"[6] uniform distribution on $[-\tau, \tau]$, which has probability density function

$$U_\tau(\eta) \equiv \begin{cases} \frac{1}{2\tau}, & -\tau \le \eta \le \tau \\ 0, & \text{otherwise} \end{cases}$$

$$\equiv \frac{1}{2\tau} 1_{[-\tau,\tau]}(\eta). \tag{5}$$

(In this example, $\pi(d\eta) = \frac{1}{2\tau} 1_{[-\tau,\tau]} d\eta$.) There are infinitely many probability distributions that assign probability 1 to $[-\tau, \tau]$; this is just one of them—one way to capture the prior information, and more. The constraint $\theta \in \Theta$ limits the *support* of π but says nothing about the probabilities π assigns to subsets of Θ. Every particular assignment—every particular choice of π—has more information than the constraint $\theta \in \Theta$, because it has information about the chance that θ is in each (measurable) subset of Θ. It expresses more than the fact that θ is an elements of Θ. Below we will show how the assumption $\theta \sim \pi$ (with π the uniform distribution) reduces the apparent uncertainty (but not the true uncertainty) in estimates of θ from Y when what we really know is just $\theta \in \Theta$.

In the linear inverse problem, the prior π is a probability distribution on Θ, which is assumed to be a measurable space. In most real inverse problems, θ is a

enormous amounts of energy finding out what would happen if the data kept pouring in." We agree. One argument made in favor of Bayesian methods is that if one does not use Bayes rule, an opponent can make "Dutch book" against him—his beliefs are not consistent in some sense. ("Dutch book" is a collection of bets that acts as a money pump: no matter what the outcome, the bettor loses money.) This is superficially appealing, but there are problems: First, beliefs have to be characterized as probabilities in the first place. Second, the argument only applies if the prior is "proper," that is, has total mass equal to one. So, the argument does not help if one wishes to use the "uninformative (uniform) prior" on an unbounded set, such as the real line. See also Eaton (2008); Eaton & Freeman (2004).

[6] Selecting a prior is not a trivial task, although most applications of Bayesian methods take the prior as given. See footnote 5. Many studies simply take the prior to be a multivariate normal distribution on a discretized set of states of the world. Some researchers tacitly invoke Laplace's Principle of Insufficient Reason to posit "uninformative" priors. Laplace's Principle of Insufficient Reason says that if there is no reason to believe that outcomes are not equally likely, assume that they are equally likely. For a discussion, see Freedman & Stark (2003). When the state of the world represents a physical quantity, some researchers use priors that are invariant under a group with respect to which the physics is invariant; similarly, when an estimation problem is invariant with respect to a group, some researchers use priors that are invariant with respect to that group. See, e.g., Lehmann & Casella (1998, p. 245ff). This is intuitively appealing and mathematically elegant, but there are difficulties. First, not all uncertainties are expressible as probabilities, so insisting on the use of any prior can be a leap. Second, why should the prior be invariant under the same group as the physics? The argument seems to be that if the physics does not distinguish between elements of an orbit of the group, the prior should not either: a change of coordinate systems should not affect the mathematical expression of our uncertainties. That is tantamount to to applying Laplace's principle of insufficient reason to orbits of the group: if there is no reason to think any element of an orbit is more likely than any other, assume that they are equally likely. But this is a fallacy because "no reason to believe that something is false" is not reason to believe that the thing is true. Absence of evidence is not evidence of absence; moreover, uncertainties have a calculus of their own, separate from the laws of physics. Third, invariance can be insufficient to determine a unique prior—and can be too restrictive to permit a prior. For example, there are infinitely many probability distributions on $\mathbb{R}^2$ that are invariant with respect to rotation about the origin, and there is no probability distribution on $\mathbb{R}^2$ that is invariant with respect to translation.

function of position—an element of an infinite-dimensional space. Constraints on θ might involve its norm or seminorm, nonnegativity or other pointwise restrictions, for example Backus (1988, 1989); Evans & Stark (2002); Parker (1994); Stark (1992b,c). There are serious technical difficulties defining probability distributions on infinite-dimensional spaces in an "uninformative" way that still captures the constraint $\theta \in \Theta$. For example, rotationally invariant priors on separable Hilbert spaces are degenerate: they either assign probability 1 to the origin, or they assign probability 1 to the event that the norm of θ is infinite—contrary to what a constraint on the norm is supposed to capture (Backus 1987).[7]

Recall that $p_\eta(y)$ is the likelihood of $\eta \in \Theta$ given y. We assume that $p_\eta(y)$ is jointly measurable with respect to η and y. The prior distribution π and the distribution $\mathbb{P}_\eta$ of the data Y given η determine the joint distribution of θ and Y. The *marginal distribution of* Y averages the distributions $\{\mathbb{P}_\eta : \eta \in \Theta\}$, weighted by π, the probability that θ is near η. The density of the marginal distribution of Y (with respect to μ) is

$$m(y) = \int_\Theta p_\eta(y)\, \pi(d\eta). \tag{6}$$

The marginal distribution of Y is also called the *predictive distribution* of Y. The information in the data Y can be used to update the prior π through Bayes' rule; the result is the *posterior distribution of* θ *given* $Y = y$:

$$\pi(d\eta|Y = y) = \frac{p_\eta(y)\ \pi(d\eta)}{m(y)}. \tag{7}$$

(The marginal density $m(y)$ can vanish; this happens with probability zero.)

In the BNM problem with uniform prior, the density of the predictive distribution of Y is

$$m(y) = \frac{1}{2\tau}\int_{-\tau}^{\tau} \phi_\eta(y)d\eta = \frac{1}{2\tau}(\Phi(y+\tau) - \Phi(y-\tau)). \tag{8}$$

The posterior distribution of θ given $Y = y$ is

$$\pi(d\eta|Y = y) = \frac{\phi_\eta(y)\frac{1}{2\tau}1_{\eta\in[-\tau,\tau]}(\eta)}{m(y)} = \frac{\phi_\eta(y)1_{\eta\in[-\tau,\tau]}(\eta)}{\Phi(y+\tau) - \Phi(y-\tau)}. \tag{9}$$

Figure 1 shows the posterior density $f_\theta(\eta|y)$ of θ (the density of $\pi(d\eta|Y = y)$) for four values of y using a uniform prior on the interval $[-3, 3]$. The posterior is a rescaled normal density with mean y, restricted to Θ. It is unimodal with mode y, but symmetric only when $y = 0$. The mode of the posterior density is the closest point in Θ to y.

In the linear inverse problem, the density of the predictive distribution is

$$m(y) = \int_\Theta f(y - K\eta)\pi(d\eta), \tag{10}$$

[7] Discretizing an infinite-dimensional inverse problem then positing a prior distribution on the discretization hides the difficulty, but does not resolve it.

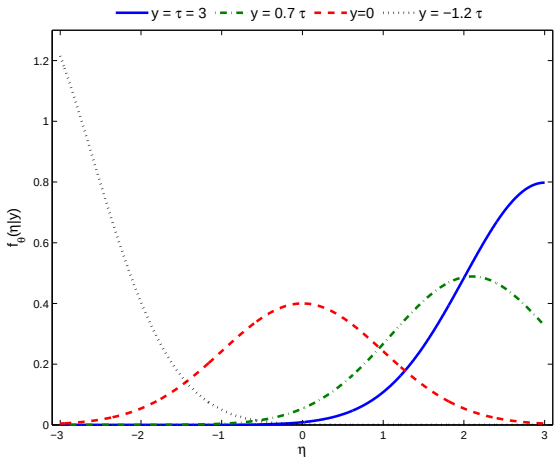

Figure 1 Posterior densities of θ for a bounded normal mean with a uniform prior on the interval $[-3, 3]$ for four values of y.

and the posterior distribution of θ given $Y = y$ is

$$\pi(\eta|Y = y) = \frac{f(y - K\eta)\pi(d\eta)}{m(y)}. \tag{11}$$

Any parameter $\lambda[\theta]$ has a *marginal posterior distribution given* $Y = y$, $\pi_\lambda(d\ell|Y = y)$, induced by the posterior distribution of θ given $Y = y$. It is defined by

$$\int_\Lambda \pi_\lambda(d\ell|Y = y) \equiv \int_{\eta:\lambda[\eta]\in\Lambda} \pi(d\eta|Y = y) \tag{12}$$

for suitably measurable sets Λ.

0.3 Estimators: What can you do with what you measure?

Estimators are mappings from the set $\mathcal{Y}$ of possible observations to some other set. Let $\mathcal{L}$ denote the set of possible values of the parameter $\lambda[\eta]$ as η ranges over Θ. *Point estimators* of $\lambda[\theta]$ assign an element $\ell \in \mathcal{L}$ to each possible observation $y \in \mathcal{Y}$. For example, in the BNM problem, $\mathcal{L} = [-\tau, \tau]$ is the set of possible values of the parameter θ, and we might consider the *truncation estimator*

$$\begin{aligned}\widehat{\theta}_\tau(y) &\equiv \begin{cases} -\tau, & y \le -\tau, \\ y, & -\tau < y < \tau \\ \tau, & y \ge \tau \end{cases} \\ &\equiv -\tau 1_{(-\infty,-\tau]}(y) + y1_{(-\tau,\tau)}(y) + \tau 1_{[\tau,\infty)}(y). \end{aligned} \tag{13}$$

We will consider other point estimators for the BNM problem below.

A *set estimator* of a parameter assigns a subset of $\mathcal{L}$ to each possible observation $y \in \mathcal{Y}$. Confidence intervals are set estimators. For example, in the BNM problem we might consider the truncated naive interval

$$\mathcal{I}_\tau(y) \equiv [-\tau, \tau] \cap [y - 1.96, y + 1.96]. \tag{14}$$

(This interval is empty if $y < -\tau - 1.96$ or $y > \tau + 1.96$.) This interval has the property that

$$\mathbb{P}_\eta\{\mathcal{I}_\tau(Y) \ni \eta\} = 0.95, \tag{15}$$

whenever $\eta \in [-\tau, \tau]$. Below we will study the *truncated Pratt interval*, a confidence interval that uses the constraint in a similar way, but tends to be shorter—especially when τ is small.

More generally, a set estimator S for the parameter $\lambda[\theta]$ that has the property

$$\mathbb{P}_\eta\{S(Y) \ni \lambda[\eta]\} \geq 1 - \alpha \tag{16}$$

for all $\eta \in \Theta$ is called a $1 - \alpha$ *confidence set for* $\lambda[\theta]$.

A $1 - \alpha$ *Bayesian credible region* for a parameter $\lambda[\theta]$ is a set that has posterior probability at least $1 - \alpha$ of containing $\lambda[\theta]$. That is, if $S_\pi(y)$ satisfies

$$\int_{S_\pi(y)} \pi_\lambda(d\ell | Y = y) \geq 1 - \alpha, \tag{17}$$

then $S_\pi(y)$ is a $1 - \alpha$ credible region. Among sets with the property (17), we may choose as the credible region the one with smallest volume: that is the smallest region that contains θ with posterior probability at least $1 - \alpha$. If the posterior distribution of $\lambda[\theta]$ has a density, that credible region is the *highest posterior density credible set*, a set of the form $S_\pi(y) = \{\eta : \pi_\lambda(d\ell | Y = y) \geq c\}$ with c chosen so that $S_\pi(y)$ satisfies (17). See, e.g., Berger (1985).

In the bounded normal mean problem with uniform prior, the highest posterior density credible region for θ is of the form

$$\{\eta \in [-\tau, \tau] : e^{-(y-\eta)^2/2} 1_{\eta \in [-\tau, \tau]}(\eta) \geq c\}, \tag{18}$$

where c depends on y and is chosen so that (17) holds. Figure 2 shows examples of the truncated naive confidence interval, the truncated Pratt confidence interval and credible regions for several values of y and τ. The truncated naive and truncated Pratt intervals are at 95% confidence level; the credible regions are at 95% credible level. When y is near 0, the truncated Pratt interval is the shortest. When $|y|$ is sufficiently large, the truncated naive interval is empty. Section 0.5 shows the confidence level of the credible region in the BNM problem.

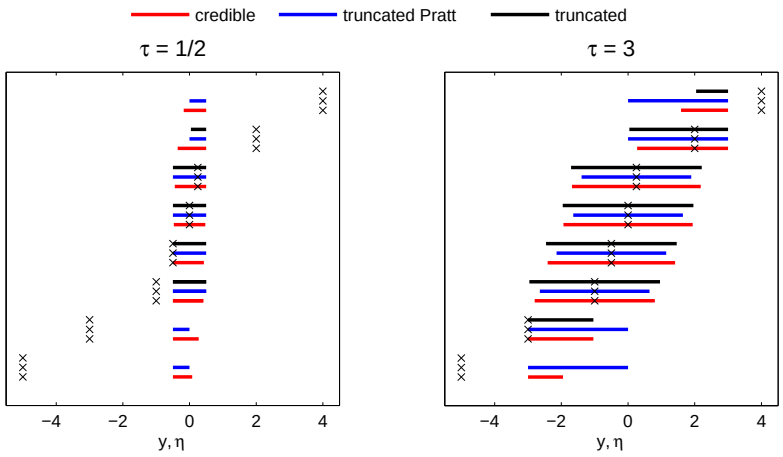

Figure 2 Examples of confidence intervals and credible sets for the BNM with $\tau = 1/2$ and $\tau = 3$, for eight different values of y. The horizontal axis represents the datum $Y = y$, indicated by an x, and the values η in in the confidence interval for the observed value of y, which are plotted as horizontal bars. The red bars are 95% maximum posterior density credible regions using a uniform prior on $[-\tau, \tau]$. The blue bars are 95% truncated Pratt intervals and the black bars are 95% truncated naive intervals. The truncated naive intervals are empty when $|y| > \tau$.

0.4 Performance of estimators: How well can you do?

This section presents several properties of estimators that can be used to define what it means for an estimator to be "good," or for one estimator to be "better" than another.

0.4.1 Bias, Variance

Statisticians commonly consider the bias and variance of point estimators. The bias is the expected difference between the parameter and the estimate of the parameter. For example, let $\widehat{\lambda}(Y)$ be an estimator of $\lambda[\theta]$. The *bias at η of $\widehat{\lambda}$* is

$$\text{bias}_\eta(\widehat{\lambda}) \equiv \mathbb{E}_\eta(\widehat{\lambda}(Y) - \lambda[\eta]) = \mathbb{E}_\eta\widehat{\lambda}(Y) - \lambda[\eta]. \tag{19}$$

when the expectation exists. The estimator $\widehat{\lambda}$ is unbiased for $\lambda[\eta]$ at η if $\text{bias}_\eta(\widehat{\lambda}) = 0$. It is *unbiased* if $\text{bias}_\eta(\widehat{\lambda}) = 0$ for all $\eta \in \Theta$. If there exists an unbiased estimator of $\lambda[\theta]$ then we say $\lambda[\theta]$ *is unbiasedly estimable* (or *U-estimable*). If $\lambda[\theta]$ is U-estimable, there is an estimator that, on average across possible samples, is equal to $\lambda[\theta]$. (Not every parameter is U-estimable.)

For example, the bias at η of the truncation estimator $\widehat{\theta}_\tau$ in the BNM problem is

$$\text{bias}_\eta(\widehat{\theta}_\tau) = \mathbb{E}_\eta \widehat{\theta}_\tau(Y) - \eta, \tag{20}$$

where

$$\begin{aligned}\mathbb{E}_\eta \widehat{\theta}_\tau(Y) &= \int_{-\infty}^{\infty} \widehat{\theta}_\tau(y)\phi_\eta(y)dy \\ &= (-\tau-\eta)\Phi(-\tau-\eta) + (\tau-\eta)\Phi(-\tau+\eta) + \\ &\quad +\phi(-\tau-\eta) - \phi(\tau-\eta),\end{aligned} \tag{21}$$

and $\phi(y) = \phi_0(y)$ is the standard Gaussian density (Eqn. (3) with $\eta = 0$) and $\Phi(y) = \int_{-\infty}^{y} \phi(x)dx$ is the standard Gaussian cumulative distribution function (cdf). Note that if $\eta = 0$, the bias at η of $\widehat{\theta}_\tau$ is zero, as one would expect from symmetry. Figure 3 shows $\text{bias}^2_\eta(\widehat{\theta}_\tau)$ as a function of η. Note that the squared bias increases as $|\eta|$ approaches τ. There, the truncation acts asymmetrically, so that the estimator is much less likely to be on one side of η than the other. For example, when $\eta = -\tau$, the estimator is never below η but often above, so the expected value of $\widehat{\theta}_\tau$ is rather larger than η—the estimator is biased.

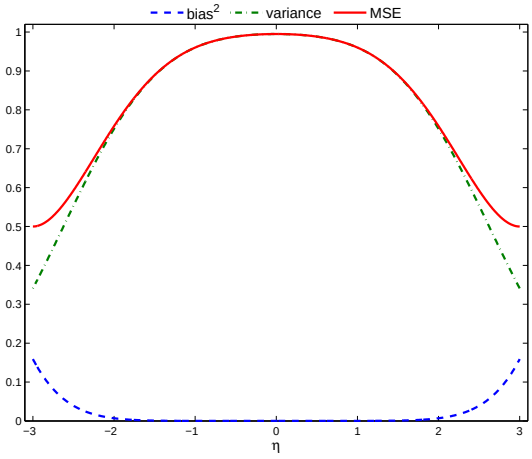

Figure 3 Squared bias, variance and MSE of the truncation estimator $\widehat{\theta}_\tau$ as a function of $\eta \in [-3, 3]$.

The *variance at* η of an estimator $\widehat{\lambda}$ of a real-valued parameter $\lambda[\theta]$ is

$$\text{Var}_\eta(\widehat{\lambda}) \equiv \mathbb{E}_\eta \left(\widehat{\lambda}(Y) - \mathbb{E}_\eta(\widehat{\lambda}(Y)\right)^2 \tag{22}$$

when the two expectations on the right exist. For example, the variance at η of the truncation estimator $\widehat{\theta}_\tau$ in the BNM problem is

$$\begin{aligned}\text{Var}_\eta \widehat{\theta}_\tau &= \int_{-\infty}^{\infty} \left(\widehat{\theta}_\tau(y) - \left(\int_{-\infty}^{\infty} \widehat{\theta}_\tau(x) \phi(x-\eta) dx \right) \right)^2 \phi(y-\eta) dy \\ &= (\tau^2 - \eta^2 - 1)\Phi(-\tau - \eta) + (\eta^2 - \tau^2 + 1)\Phi(\tau - \eta) + (\eta - \tau)\phi(\tau + \eta) \\ &\quad -(\tau + \eta)\phi(\tau - \eta) + \tau^2 - (\, \mathbb{E}_\eta \widehat{\theta}_\tau(Y)\,)^2. \end{aligned} \tag{23}$$

Figure 3 shows $\text{bias}_\eta^2(\widehat{\theta}_\tau)$ and $\text{Var}_\eta(\widehat{\theta}_\tau)$ as functions of η for $\tau = 3$. The variance is biggest near the origin, where truncation helps least—if τ is large, the variance at $\eta = 0$ approaches the variance of the additive Normal noise, $\sigma^2 = 1$. Near $\pm\tau$, the truncation reduces the variance because it keeps the estimator from being very far from η on one side. But as we have seen, that introduces bias.

Neither bias nor variance alone suffices to quantify the error of an estimator: even if an estimator is unbiased, its variance can be so large that it is rarely close to the parameter it estimates.[8] Conversely, an estimator can have small or zero variance, and yet its bias can be so large that it is never close to the parameter. Small variance means small scatter; small bias means that on average the estimator is right.[9] Insisting that estimators be unbiased can be counter-productive, for a number of reasons. For instance, there are problems in which there is no estimator that is unbiased for all $\eta \in \Theta$. And allowing some bias can reduce the overall MSE, or another measure of the accuracy of an estimator. We would agree that it is important to have a bound on the magnitude of the bias—but that bound need not be zero.

The next subsection presents a general approach to defining the quality of estimators.

0.4.2 Loss and Risk

A more general way to evaluate and compare estimators is through a *loss function* that measures the loss we incur if we take a particular *action* ℓ when the true value of the parameter is $\lambda[\eta]$. The *risk at* η of an estimator is the expected value of the loss when that estimator is used and the world is in state η:

$$\rho_\eta(\widehat{\lambda}, \lambda[\eta]) \equiv \mathbb{E}_\eta \, \text{loss}(\widehat{\lambda}(Y), \lambda[\eta]). \tag{24}$$

For example, consider estimating a real-valued parameter $\lambda[\theta]$ from data Y. For point estimates, actions are possible parameter values—real numbers in this case. We

[8] Three statisticians go deer hunting. They spot a deer. The first shoots; the shot goes a yard to the left of the deer. The second shoots; the shot goes a yard to the right of the deer. The third statistician shouts, "we got him!"

[9] Estimation is like shooting a rifle. If the rifle is well made, the shots will hit nearly the same mark; but if its sights are not adjusted well, that mark could be far from the bullseye. Conversely, shots from a poorly made rifle will be scattered, no matter how well adjusted its sights are. The quality of the rifle is analogous to the variance: small variance corresponds to high quality—low scatter. The adjustment of the sights is analogous to the bias: low bias corresponds to well adjusted sights—on average, the estimate and the shots land where they should.

could define the loss for taking the action $\ell \in \mathbb{R}$ when the true parameter value is $\lambda[\eta]$ to be $\text{loss}(\ell, \lambda[\eta]) = |\ell - \lambda[\eta]|^p$, for example. Scientific context should dictate the loss function,[10] but most theoretical work on point estimation uses *squared-error loss*: $\text{loss}(\ell, \lambda[\eta]) = |\ell - \lambda[\eta]|^2$. For squared-error loss, the risk at η of the estimator $\hat{\lambda}$ is the *mean squared error*:

$$\text{MSE}_\eta(\widehat{\lambda}) \equiv \mathbb{E}_\eta |\widehat{\lambda}(Y) - \lambda[\eta]|^2, \tag{25}$$

when the expectation exists. The MSE of a point estimate can be written as the square of the bias plus the variance:

$$\text{MSE}_\eta(\widehat{\lambda}) = \text{bias}_\eta^2(\widehat{\lambda}) + \text{Var}_\eta(\widehat{\lambda}). \tag{26}$$

In the BNM problem, the MSE at η of the truncation estimator $\widehat{\theta}_\tau$ of θ is

$$\text{MSE}_\eta(\widehat{\theta}) = \mathbb{E}_\eta |\widehat{\theta} - \eta|^2 = \text{bias}_\eta^2 \widehat{\theta}_\tau + \text{Var}_\eta(\widehat{\theta}_\tau). \tag{27}$$

Figure 3 plots $\text{MSE}_\eta(\widehat{\theta})$ as a function of η for $\tau = 3$. Near the origin the MSE is dominated by the variance; near the boundaries the squared bias becomes important, too.

For set estimators, one measure of loss is the size (Lebesgue measure) of the set; the risk at η is then the expected measure of the set when the true state of the world is η. For example, let $\mathcal{I}$ be a (suitably measurable) set of confidence set estimators—estimators that, with probability at least $1 - \alpha$ produce a (Lebesgue-measurable) set that contains $\lambda[\eta]$ when η is the true state of the world. That is, if $I \in \mathcal{I}$ then

$$\mathbb{P}_\eta\{I(Y) \ni \lambda[\eta]\} \geq 1 - \alpha \tag{28}$$

for all $\eta \in \Theta$. If $I(Y)$ is always Lebesgue-measurable,

$$\rho_\eta(I, \lambda[\eta]) \equiv \mathbb{E}_\eta \int_\ell I(Y) d\ell \tag{29}$$

is the expected measure of the set, a reasonable risk function for set estimates; see Evans et al. (2005). For example, Figure 4 shows the expected length of the truncated naive interval (and some other intervals) for the bounded normal mean, as a function of η. We will discuss this plot below.

0.4.3 Decision theory

Decision theory provides a framework for selecting estimators. Generally, the risk of an estimator $\widehat{\lambda}$ of a parameter $\lambda[\theta]$ depends on the state θ that the world is in:

[10] For example, consider the problem of estimating how deep to set the pilings for a bridge. If the pilings are set too deep, the bridge will cost more than necessary. But if the pilings are set too shallow, the bridge will collapse, and lives could be lost. Such a loss function is asymmetric and highly nonlinear. Some theoretical results in decision theory depend on details of the loss function (for instance, differentiability, convexity, or symmetry), and squared error is used frequently because it is analytically tractable.

some estimators do better when θ has one value; some do better when θ takes a different value. Some estimators do best when θ is a random variable with a known distribution π. If we knew θ (or π), we could pick the estimator that had the smallest risk. But if we knew θ, we would not need to *estimate* $\lambda[\theta]$: we could just *calculate* it.

Decision theory casts estimation as a two-player game: Nature versus statistician. Frequentists and Bayesians play similar—but not identical—games. In both games, Nature and the statistician know Θ; they know how data will be generated from θ; they know how to calculate $\lambda[\theta]$ from any $\theta \in \Theta$; and they know the loss function ρ that will be used to determine the payoff of the game. Nature selects θ from Θ without knowing what estimator $\widehat{\lambda}$ the statistician plans to use. The statistician, ignorant of θ, selects an estimator $\widehat{\lambda}$. The referee repeatedly generates Y from the θ that Nature selected and calculates $\widehat{\lambda}(Y)$. The statistician has to pay the average value of loss$(\widehat{\lambda}(Y), \lambda[\theta])$ over all those values of Y; that is, $\rho_\theta(\widehat{\lambda}, \lambda[\theta])$.

The difference between the Bayesian and frequentist views is in how Nature selects θ from Θ. Bayesians suppose that Nature selects θ at random according to the prior distribution π—and that the statistician knows *which* distribution π Nature is using.[11] Frequentists do not suppose that Nature draws θ from π, or else they do not claim to know π. In particular, frequentists do not rule out the possibility that Nature might select θ from Θ to maximize the amount the statistician has to pay on average. The difference between Bayesians and frequentists is that Bayesians claim to know more than frequentists: both incorporate constraints into their estimates and inferences, but Bayesians model uncertainty about the state of the world using a particular probability distribution in situations where frequentists would not. Bayesians speak of the probability that the state of the world is in a given subset of Θ. For frequentists, such statements generally do not make sense: the state of the world is not random with a known probability distribution—it is simply unknown. The difference in the amount of information in "θ is a random element of Θ drawn from a known probability distribution" and "θ is an unknown element of Θ" can be small or large, depending on the problem and the probability distribution.

So, which estimator should the statistician use? The next two subsections show those differences in the rules of the game lead to different strategies Bayesians and frequentists use to select estimators.

Minimax Estimation

One common strategy for frequentists is to minimize the maximum they might be required to pay. That is, to use the estimator (among those in a suitable class) that has the smallest worst-case risk over all possible states of the world $\eta \in \Theta$.

[11] Some Bayesians do not claim to know π, but claim to know that π itself was drawn from a probability distribution on probability distributions. Such *hierarchical priors* do not really add any generality: they are just a more complicated way of specifying a probability distribution on states of the world. A weighted mixture of priors is a prior.

For example, consider the BNM problem with MSE risk. Restrict attention to the class of estimators of θ that are affine functions of the datum Y; that is, the set of estimators $\widehat{\theta}$ of the form

$$\widehat{\theta}_{ab}(y) = a + by, \tag{30}$$

where a and b are real numbers. We can choose a and b to minimize the maximum MSE for $\eta \in \Theta$. The resulting estimator, $\widehat{\theta}_A$, is the *minimax affine estimator* (for MSE loss). To find $\widehat{\theta}_A$ we first calculate the MSE of $\widehat{\theta}_{ab}$. Let $Z \sim N(0,1)$ be a standard Gaussian random variable, so that $Y \sim \eta + Z$ if $\theta = \eta$.

$$\begin{aligned} MSE_\eta(\widehat{\theta}_{ab}) &= \mathbb{E}_\eta(\widehat{\theta}_{ab}(Y) - \eta)^2 \\ &= \mathbb{E}_\eta(a + b(\eta + Z) - \eta)^2 \\ &= \mathbb{E}_\eta(a + (b-1)\eta + bZ)^2 \\ &= (a + (b-1)\eta)^2 + b^2 \\ &= a^2 + 2a(b-1)\eta + (b-1)^2\eta^2 + b^2. \end{aligned} \tag{31}$$

This is quadratic in η with positive leading coefficient, so the maximum will occur either at $\eta = -\tau$ or $\eta = \tau$, whichever has the same sign as $a(b-1)$. Since the leading term is nonnegative and the second term will be positive if $a \neq 0$, it follows that $a = 0$ for the minimax affine estimator. We just have to find the optimal value of b. The MSE of the minimax affine estimator is thus $\tau^2(b-1)^2 + b^2$, a quadratic in b with positive leading coefficient. So the minimum MSE will occur at a stationary point with respect to b, and we find

$$2\tau^2(b-1) + 2b = 0, \tag{32}$$

i.e., the minimax affine estimator is

$$\widehat{\theta}_A(y) = \frac{\tau^2}{\tau^2 + 1} \cdot y. \tag{33}$$

This is an example of a *shrinkage estimator*: it takes the observation and shrinks it towards the origin, since $\tau^2/(\tau^2+1) \in (0,1)$. Multiplying the datum by a number less than 1 makes the variance of the estimator less than the variance of Y—the variance of the minimax affine estimator is $\tau^2/(\tau^2+1)^2$, while the variance of Y is one. Shrinkage also introduces bias: the bias is $\eta(\tau^2/(\tau^2+1) - 1)$. But because we know that $\eta \in [-\tau, \tau]$, the square of bias is at most $\tau^2(\tau^2/(\tau^2+1) - 1)^2$, and so the MSE of the minimax affine estimator is at most

$$\frac{\tau^2}{(\tau^2+1)^2} + \tau^2\left(1 - \frac{\tau^2}{\tau^2+1}\right)^2 = \frac{\tau^2}{\tau^2+1} < 1. \tag{34}$$

By allowing some bias, the variance can be reduced enough that the MSE is smaller than the MSE of the raw estimator Y of θ (the MSE of Y as an estimator of θ is 1 for all $\eta \in \Theta$)

Affine estimators use the data in a very simple way. The truncation estimator $\widehat{\theta}_\tau$ is not affine. What about even more complicated estimators? If we allowed an estimator to depend on Y in an arbitrary (but measurable) way, how small could its maximum MSE be than the maximum MSE of the minimax affine estimator?

In the BNM problem, the answer is "not much smaller:" the maximum MSE of the minimax (nonlinear) estimator of θ is no less than $4/5$ the maximum MSE of the minimax affine estimator of θ (Donoho at al. 1990). Let $a \wedge b \equiv \max(a, b)$ and $a \vee b \equiv \min(a, b)$. Figure 6 also shows the risk of the *truncated minimax affine estimator*

$$\widehat{\theta}_{A\tau}(y) = \tau \wedge (-\tau \vee \tau(\tau^2 + 1)^{-1} y) \tag{35}$$

as a function of η. This estimator improves the minimax affine estimator by ensuring that the estimate is in $[-\tau, \tau]$ even when $|Y|$ is very large. For $\tau = 1/2$, the maximum risk of the truncated minimax affine estimator is about 25% larger than the lower bound on the maximum risk of any nonlinear estimator. For $\tau = 3$, its maximum risk is about 12% larger than the lower bound.

We can also find confidence set estimators in various classes that are minimax for a risk function, such as expected length. Stark (1992a) studies the minimax length of fixed-length confidence intervals for a BNM when the intervals are centered at an affine function of Y. Zeytinoglu & Mintz (1984) study the minimax length of fixed-length confidence intervals for a BNM when the intervals are centered at nonlinear functions of Y. Evans et al. (2005) study minimax expected size confidence sets where the size can vary with Y. They show that when $\tau \leq 2\Phi^{-1}(1 - \alpha)$, the minimax expected size confidence interval for the BNM problem is the *truncated Pratt interval*:

$$\mathcal{I}_{\mathrm{P}\tau}(Y) \equiv \mathcal{I}_P(Y) \cap [-\tau, \tau], \tag{36}$$

where $\mathcal{I}_P(Y)$ is the Pratt interval (Pratt 1961):

$$\mathcal{I}_{\mathrm{P}}(Y) \equiv \begin{cases} [(Y - \Phi^{-1}(1-\alpha)), 0 \vee (Y + \Phi^{-1}(1-\alpha))], & Y \leq 0 \\ [0 \wedge (Y - \Phi^{-1}(1-\alpha)), Y + \Phi^{-1}(1-\alpha)], & Y > 0. \end{cases} \tag{37}$$

The truncated Pratt interval $\mathcal{I}_{\mathrm{P}\tau}$ has minimax expected length among all $1 - \alpha$ confidence intervals for the BNM when $\tau \leq 2\Phi^{-1}(1 - \alpha)$. (For $\alpha = 0.05$, that is $\tau \leq 3.29$.) For larger values of τ, the minimax expected length confidence procedure can be approximated numerically; see Evans et al. (2005); Schafer & Stark (2009). Figure 4 compares the expected length of the minimax expected length confidence interval for a BNM with the expected length of the truncated naive interval as a function of η. It also shows the expected length of the Bayes credible region using a uniform prior (see the next section). Note that the minimax expected length interval has the smallest expected length when η is near zero.

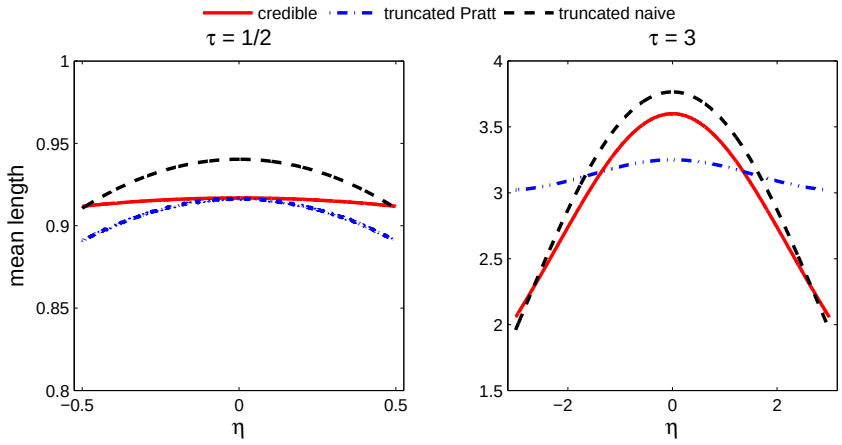

Figure 4 Expected length of the 95% credible, truncated naive and truncated Pratt interval for the BNM as a function of η for $\tau = 1/2$ and $\tau = 3$. For reference, the length of the naive confidence interval, which does not use the information $\theta \in \Theta$, is $2 \times 1.96 = 3.92$ for all y, so its expected length is 3.92 for all η.

Bayes risk

The average ρ-risk of an estimator for prior π is the mean risk of the estimator when the parameter θ is chosen from the prior π:

$$\rho(\widehat{\lambda}, \lambda; \pi) \equiv \int_{\eta} \rho_{\eta}(\widehat{\lambda}, \lambda[\eta])\pi(d\eta). \tag{38}$$

The Bayes ρ-risk for prior π is the smallest average ρ-risk of any estimator for prior π:

$$\rho(\lambda; \pi) \equiv \inf_{\widehat{\lambda}} \rho(\widehat{\lambda}, \lambda; \pi), \tag{39}$$

where the infimum is over some suitable class of estimators. An estimator whose average ρ-risk for prior π is equal to the Bayes ρ-risk for prior π is called a *Bayes estimator*.

Consider estimating a real-valued parameter $\lambda[\theta]$ under MSE risk. It is well known that the Bayes estimator is the mean of the marginal posterior distribution of $\lambda[\theta]$ given Y, $\pi_{\lambda}(d\ell|Y = y)$:

$$\hat{\lambda}_{\pi}(Y) \equiv \int \ell \pi_{\lambda}(d\ell|Y = y). \tag{40}$$

See, e.g., Berger (1985); Lehmann & Casella (1998).[12]

[12]This follows from the general result that if X is a real-valued random variable with probability measure μ, then among all constants c, the smallest value of $\int (x - c)^2 \mu(dx)$ is attained when $c = \int x\mu(dx) = \mathbb{E}(X)$. Let $\mu(dx) = \pi_{\lambda}(dx|Y = y)$.

For example, in the BNM problem the average MSE risk of the estimator $\widehat{\lambda}(Y) = Y$ of θ for the uniform prior on $[-\tau, \tau]$ is

$$\rho(Y, \theta; \pi) \equiv \frac{1}{2\pi} \int_{-\tau}^{\tau} \mathbb{E}_\eta (Y - \eta)^2 d\eta = 1. \tag{41}$$

The Bayes risk is

$$\rho(\theta, \pi) = \inf_{\widehat{\lambda}} \rho(\widehat{\lambda}, \theta; \pi). \tag{42}$$

And the Bayes estimator is

$$\hat{\theta}_\pi(Y) \equiv Y - \frac{\phi(\tau - Y) - \phi(\tau + Y)}{\Phi(\tau - Y) - \Phi(-\tau - Y)}. \tag{43}$$

Because the posterior distribution of θ has a unimodal density, a level set of the posterior density of the form $S(Y) = \{\eta : \pi(d\eta | Y = y) \geq c\}$ is an interval. If c is chosen so that the posterior probability of $S(Y)$ is $1 - \alpha$, S is a $1 - \alpha$ credible interval. Figure 4 shows the expected length of the Bayes credible region for the BNM using a uniform prior.

Bayes/Minimax duality

The Bayes risk depends on the parameter to be estimated and the loss function—and also on the prior. Consider a set of priors that includes point masses at each element of Θ (or at least, priors that can assign probability close to 1 to small neighborhoods of each $\eta \in \Theta$). Vary the prior over that set to find a prior for which the Bayes risk is largest. Such priors are called "least favorable."

The frequentist minimax problem finds an $\eta \in \Theta$ for which estimating $\lambda[\eta]$ is hardest. The least favorable prior is a distribution on Θ for which estimating $\lambda[\eta]$ is hardest *on average* when θ is drawn at random from that prior. Since the set of priors in the optimization problem includes distributions concentrated near the $\eta \in \Theta$ for which the minimax risk is attained, the Bayes risk for the least favorable prior is no smaller than the minimax risk. Perhaps surprisingly, the Bayes risk for the least favorable prior is in fact equal to the minimax risk under mild conditions. See, for example, Berger (1985); Lehmann & Casella (1998).[13]

Because the Bayes risk for the least favorable prior is a lower bound on the minimax risk, comparing the Bayes risk for any particular prior with the minimax risk measures how much information the prior adds: whenever the Bayes risk is smaller than the minimax risk, it is because the prior is adding more information than simply the constraint $\theta \in \Theta$ or the data. When that occurs, it is prudent to ask

[13] Casella & Strawderman (1981) derive the (nonlinear) minimax MSE estimator for the BNM problem for small τ by showing that a particular 2-point prior is least favorable when τ is suffi ciently small and that a 3-point prior is least favorable when τ is a little larger. They show that the number of points of support of the least favorable prior grows as τ grows. Similarly, Evans et al. (2005) construct minimax expected measure confi dence intervals by characterizing the least favorable prior distributions; see also Schafer & Stark (2009).

whether the statistician *knows* that θ is drawn from the distribution π, or has adopted π to capture the constraint that $\theta \in \Theta$. If the latter, the Bayes risk understates the true uncertainty.

0.5 Frequentist performance of Bayes estimators for a BNM

In this section, we examine Bayes estimates in the BNM problem with uniform prior from a frequentist perspective.[14] In particular, we look at the maximum MSE of the Bayes estimator of θ for MSE risk, and at the frequentist coverage probability and expected length of the 95% Bayes credible interval for θ.[15]

0.5.1 MSE of the Bayes Estimator for BNM

Figure 5 shows the squared bias, variance and MSE of the Bayes estimator of a bounded normal mean for MSE risk using a uniform prior, for $\tau = 3$. The figure can be compared with figure 3, which plots the squared bias, variance and MSE of the truncation estimator for the same problem. Note that the squared bias of the Bayes estimate is comparatively larger, and the variance is smaller. The Bayes estimate has its largest MSE when $|\eta|$ is close to τ, which is where the truncation estimate has its smallest MSE. Figure 6 compares the MSE of the Bayes estimator of θ with the MSE of the truncation estimator $\widehat{\theta}_\tau$ and the minimax affine estimator $\widehat{\theta}_A$, as a function of θ, for $\tau = 1/2$ and $\tau = 3$. When $\tau = 1/2$, the truncation estimator is *dominated* by both of the others: the other two have risk functions that are smaller for every $\eta \in \Theta$. The risk of the Bayes estimator is smaller than that of the nonlinear minimax estimator except when $|\eta|$ is close to τ, although the two are quite close everywhere. When $\tau = 3$, none of the estimators dominates either of the others: there are regions of $\eta \in \Theta$ where each of the three has the smallest risk. The truncation estimator does best when $|\eta|$ is close to τ and worst with $|\eta|$ is near zero. The Bayes estimator generally has the smallest risk of the three, except when $|\eta|$ is close to τ, where its risk is much larger than that of the truncation estimator and noticeably larger than that of the minimax estimator.

0.5.2 Frequentist coverage of the Bayesian Credible Regions for BNM

Figure 7 plots the coverage probability of the 95% Bayes maximum posterior density credible region as a function of $\eta \in \Theta$, for $\tau = 1/2$ and $\tau = 3$. For $\tau = 1/2$, the

[14] Generally, Bayes point estimators are biased. For more about (frequentist) risk functions of Bayes estimators, see, e.g., Lehmann & Casella (1998, p. 241ff).

[15] For a discussion of the frequentist consistency of Bayes estimates, see Diaconis & Freedman (1986, 1998).

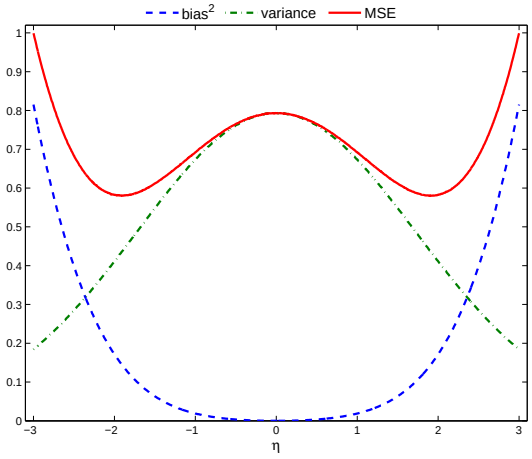

Figure 5 Squared bias, variance and MSE of the Bayes estimator of a bounded normal mean using a uniform prior on $[-3, 3]$.

coverage probability is nearly 100% when η is in the middle of Θ, but drops precipitously as η approaches $\pm\tau$, where it is 68%. For $\tau = 3$, the coverage probability is smallest when η is near zero, where it is 90.9%. For most values of η, the coverage probability of the credible region is at least 95%. On average over $\eta \in \Theta$, the coverage probability is about right, but for some values is is far too large and for some it is far too small.

0.5.3 Expected length of the Bayesian Credible Region for BNM

The frequentist coverage probability of the Bayesian 95% credible region is close to 95% except when θ is near $\pm\tau$. Does the loss in coverage probability buy a substantial decrease in the expected length of the interval? Figure 4 shows that that is not the case: when τ is small, the expected length truncated Pratt interval is nowhere longer than that of the maximum posterior density credible region; when τ is moderate, the truncated naive interval has expected length only slightly larger than that of the credible region, and the truncated Pratt interval has smaller expected length for much of Θ.

0.6 Summary

Inverse problems involve estimating or drawing inferences about a *parameter*—a property of the unknown state of the world—from indirect, noisy data. Prior information about the state of the world can reduce the uncertainty in estimates and inferences. In physical science, prior information generally consists of constraints.

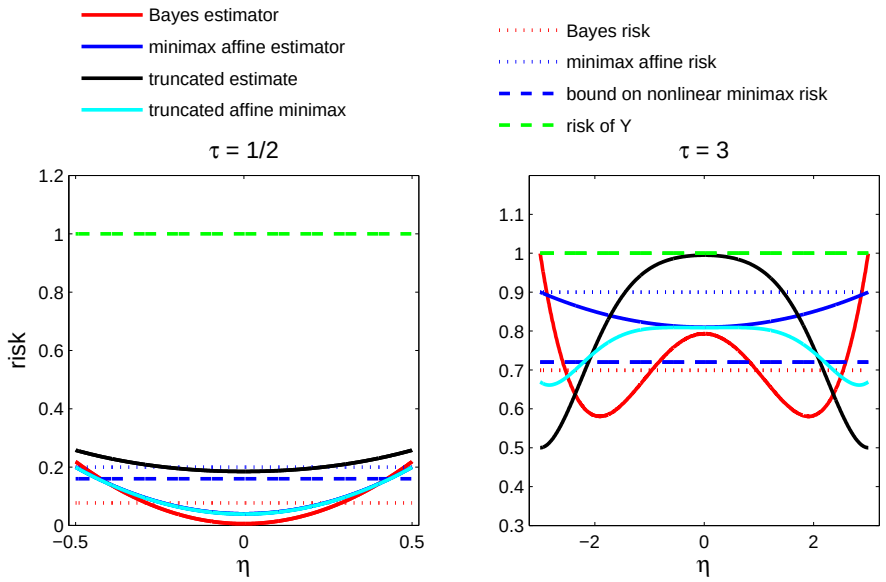

Figure 6 MSE risk of the naive estimator Y, the truncation estimator, the Bayes (for uniform prior), the minimax affine estimator and the truncated minimax affine estimator for the BNM problem. For each estimator, the risk at η is plotted as a function of η for $\tau = 1/2$ and $\tau = 3$. The risk of Y is constant and equal to 1. The other three horizontal lines are the Bayes risk of the Bayes estimator, the minimax risk of the minimax affine estimator and a lower bound on the minimax risk of any nonlinear estimator. Some of the risks are computed analytically; others using 6×10^6 simulations. The fluctuations in the empirical approximation are less than the linewidth in the plots.

For example, mass density is nonnegative; velocities are bounded; energies are finite. Frequentist methods can use such constraints directly. Bayesian methods require that constraints be re-formulated as prior probability distribution. That re-formulation inevitably adds information not present in the original constraint.

Quantities that can be computed from the data (without knowing the state of the world or the value of the parameter) are called *estimators*. Point estimators map the data into possible values of the parameter. The sample mean is an example of a point estimator. Set estimators map the data into sets of possible values of the parameter. Confidence intervals are examples of set estimators. There are also other kinds of estimators. And within types of estimators, there are classes with different kinds of functional dependence on the data. For example, once might consider point estimators that are affine functions of the data, or that are arbitrary nonlinear functions of the data.

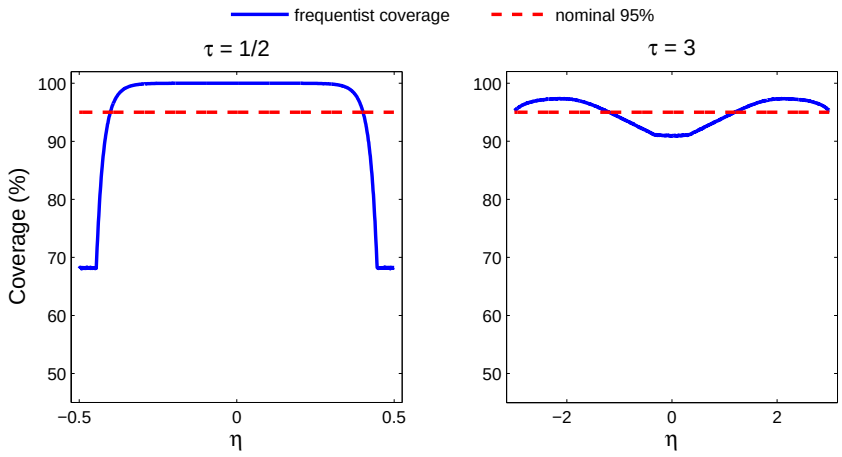

Figure 7 Frequentist coverage as a function of η of the 95% credible interval for the BNM with a uniform prior on $[-\tau, \tau]$ for $\tau = 1/2$ and $\tau = 3$.

Estimators can be compared using *risk functions*. Mean squared error is a common risk function for point estimators. Expected measure of a confidence set is a reasonable risk function for set estimators. Scientific considerations should enter the choice of risk functions, although risk functions are often selected for their mathematical tractability rather than their relevance.

The risk depends on the true state of the world. Typically, no estimator has smallest risk for all possible states of the world. Decision theory makes the risk tradeoff explicit. For example, one might consider the worst-case risk over all possible states of the world. The estimator with the smallest worst-case risk is the *minimax estimator*, a frequentist notion. Or one might consider the average risk if the world were drawn at random from a given prior probability distribution. The estimator with the smallest average risk is the *Bayes estimator* for that prior distribution. If the state of the world is not random, or random but with unknown probability distribution, frequentist methods may be preferable to Bayesian methods.

Bayesian methods can be evaluated from a frequentist perspective, and vice versa. For example, one can calculate the Bayes risk of a frequentist estimator, or the maximum risk of a Bayes estimator. The frequentist notion of a confidence set is similar to the Bayesian notion of credible region, but the two are not identical. The frequentist coverage probability of a Bayesian $1 - \alpha$ credible set is typically not $1 - \alpha$ for all parameter values: it can be much higher for some and much lower for others. There is a duality between Bayesian and frequentist methods: under mild technical conditions, the risk of the Bayes estimator for the *least favorable prior* is equal to the risk of the minimax estimator. When the Bayes risk for a given prior is less than the minimax risk, the apparent uncertainty in the Bayes estimate has

been reduced by the choice of prior: the prior adds information not present in the constraints.

Bibliography

Backus GE 1987 Isotropic probability measures in infinite-dimensional spaces. *Proc. Natl. Acad. Sci.* **84**, 8755–8757.

Backus GE 1988 Comparing hard and soft prior bounds in geophysical inverse problems. *Geophys. J.* **94**, 249–261.

Backus GE 1989 Confidence set inference with a prior quadratic bound. *Geophys. J.* **97**, 119–150.

Berger JO 1985 *Statistical Decision Theory and Bayesian Analysis. 2nd Edn*. Springer-Verlag.

Casella G and Strawderman WE 1981 Estimating a bounded normal mean. *Ann. Stat.* **9**, 870–878.

Diaconis P and Freedman DA 1986 On the consistency of Bayes estimates. *Ann. Stat.* **14**, 1–26.

Diaconis P and Freedman DA 1998 Consistency of Bayes estimates for nonparametric regression: normal theory. *Bernoulli* **4**, 411–444.

Donoho DL 1995 Statistical estimation and optimal recovery. *Ann. Stat.* **22**, 238–270.

Donoho DL, Liu RC and MacGibbon B 1990 Minimax risk over hyperrectangles, and implications. *Ann. Stat.* **18**, 1416–1437.

Eaton ML 2008 Dutch book in simple multivariate normal prediction: another look. In D Nolan and T Speed, editors. *Probability and Statistics: Essays in Honor of David A. Freedman*. Institute of Mathematical Statistics.

Eaton ML and Freedman DA 2004 Dutch book against some objective priors. *Bernoulli* **10**, 861–872.

Evans SN, Hansen B and Stark PB 2005 Minimax expected measure confidence sets for restricted location parameters. *Bernoulli* **11**, 571–590.

Evans SN and Stark PB 2002 Inverse problems as statistics. *Inverse Problems* **18**, R1–R43.

Freedman DA and Stark PB 2003 What is the Chance of an Earthquake? in *NATO Science Series IV: Earth and Environmental Sciences* **32**, pp.201–213.

Freedman DA 1995 Some Issues in the Foundations of Statistics. *Foundations of Science* **1**, 19–39.

Le Cam L 1986 Asymptotic Methods in Statistical Decision Theory. Springer-Verlag, New York.

Lehmann EL and Casella G 1998 *Theory of Point Estimation 2nd Edn*. Springer-Verlag.

Parker RL 1994 *Geophysical Inverse Theory*. Princeton University Press.

Pratt JW 1961 Length of confidence intervals. *J. Am. Stat. Assoc.* **56**, 549–567.

Schafer CM and Stark PB 2009 Constructing confidence sets of optimal expected size. *J. Am. Stat. Assoc.* (in press).

Stark PB 1992a Affine minimax confidence intervals for a bounded normal mean. *Stat. Probab. Lett.* **13**, 39–44.

Stark PB 1992b Inference in infinite-dimensional inverse problems: Discretization and duality. *J. Geophys. Res.* **97**, 14,055–14,082.

Stark PB 1992c Minimax confidence intervals in geomagnetism. *Geophys. J. Intl.* **108**, 329–338.

Zeytinoglu M and Mintz M 1984 Optimal fixed size confidence procedures for a restricted parameter space. *Ann. Stat.* **12**, 945–957.